# SPEECH DAT CAR. A Large Speech Database For Automotive Environments

**Asunción Moreno (1), Borge Lindberg (2), Christoph Draxler (3), Gael Richard (4),**

**Khalid Choukry (5), Stephan Euler (6), Jeff Allen (5)**

(1) Universidad Politécnica de Cataluña, Spain; (2) CPK, Denmark (3) IPSK of the University of Munich.
(4)Lernout & Hauspie , France (5), ELRA, France, (6), Bosch Germany,

Address: (1) UPC, Jordi Girona 1-3, 08034 Barcelona Spain
e-mail asuncion@gps.tsc.upc.es

**Abstract**

SpeechDat-Car aims to develop a set of speech databases to support training and testing of multilingual speech recognition applications in the car environment. As a result, a total of ten (10) equivalent and similar resources will be created. The 10 languages are Danish, Britsh English, Finnish, Flemish/Dutch, French, German, Greek, Italian, Spanish and American English. For each language 600 sessions will be recorded (from at least 300 speakers) in seven characteristic environments (low speed, high speed with audio equipment on, etc.). This paper gives an overview of the project with a focus on production phases (speaker recruitment, recording platforms, annotation….)

.

## 1. Introduction

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 new in-car applications to develop while taking care of safety aspects. 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 (Langmann (1998)). The European project SpeechDat-Car[1] aims at providing a set of uniform, coherent databases for nine European languages and for American English.

SpeechDat-Car continues the success of the SpeechDat project [Draxler (1998); Höge (1998); Höge(1999)] in developing large-scale speech resources for a wide range of languages and for in-car applications (voice dialling, car accessories control, etc.). It will produce resources for ten languages: Danish, English, Finnish, Flemish/Dutch, French, German, Greek, Italian, Spanish, and American English. 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, Nokia, Sonofon, Tawido, Vocalis), and universities (CPK, Denmark; DMI, Finland; IPSK, Germany; IRST, Italy; SPEX, Netherlands; UPC, Spain; WCL, Greece).

It is also important to note that SpeechDat-Car commits itself to a strict validation protocol to ensure optimal quality and exchangeability of the databases (van den Heuvel (1999)).

More precisely, this paper gives an overview of the project with a focus on production phases. It is organised as follows: the next section describes the database specifications (database content, recording platforms and validation procedures). Then, Section 3 provides additional information on speakers characteristics and an extensive description of the annotation procedure and tools is given in section 4. The paper is then concluded with a short section about database availability and dissemination.

## 2. Database specifications

Each database produced in the SpeechDat car project is intended to provide enough data to train speaker independent recognition systems. Database contents were designed to cope with different applications. The design includes a phonetically balanced corpus to train basic speech recognition systems and an application corpus. The application corpus was addressed to two applications: Telecommunication systems (IVR, dialling, remote access to teleservices and servers) and Car equipment (radiotelephones, car radio, car accessories, navigation).

Every database comprises recordings of 600 different sessions from at least 300 speakers. A session consist of either 119 or 129 read and spontaneous items recorded by five microphones installed in a car. Signals from four of these microphones are recorded and stored in a mobile platform installed in the car. The signal from the fifth microphone is transmitted simultaneously by GSM to a fixed platform connected to an ISDN telephone interface. This signal is recorded in A-law format.

Recordings are made in different recording conditions. There are defined 7 environment conditions. Every environment is equally represented in the final database:

- car stopped by motor running
- car in town traffic
- car in town traffic, with noisy conditions
- car moving at a low speed with rough road conditions
- car moving at a low speed with rough road conditions, with noisy conditions
- car moving at a high speed with good road conditions
- car moving at a high speed with good road conditions with audio equipment on

[1] SpeechDat-Car started in April 1998 in the 4th EC framework under project code LE4-8334 with a 30 months' project duration.

| Items | Corpus contents |
|---|---|
| Digits and strings of digits | |
| 1 | sequence of 10 isolated digits |
| 1 | sheet number (4+ digits) |
| 1 | spontaneous telephone number |
| 3 | read telephone numbers |
| 1 | credit card number (16 digits) |
| 1 | PIN code (6 digits) |
| 4 | isolated digits |
| Dates | |
| 1 | spontaneous date, e.g. birthday |
| 1 | prompted date, word style |
| 1 | relative and general date exp. |
| Spellings | |
| 1 | spontaneous, e.g. own forename |
| 1 | spelling of direct. city name |
| 4 | real word/name |
| 1 | artificial name for coverage |
| Money amount/ natural number | |
| 1 | money amount |
| 1 | natural number |
| Names | |
| 1 | spontaneous, e.g. own forename |
| 1 | city of growing up (spontaneous) |
| 2 | most frequent cities |
| 2 | company/agency /street names |
| 1 | forename/surname |
| Times | |
| 1 | time of day (spontaneous) |
| 1 | time phrase (word style) |
| Application words | |
| 13 | Mobile phone Application words |
| 22 | IVR functions keywords |
| 32 | car products keywords |
| 2 | voice activation keywords |
| 2 | language dependent keywords |
| Phonetically rich words | |
| 4 | phonetically rich words |
| Sentences | |
| 2 | phrases using an application word |
| 9 | phonetically rich sentences |
| 10 | Prompts for spontaneous speech |

TABLE 1. SpeechDat Car Database Contents

Each session is manually annotated. Only speech recorded by the close talk microphone is annotated The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting acoustic events of the utterance. Extra marks point to mispronuntiation, truncations, uninteligible words and extra noises.

A rigorous validation procedure is applied to each database produced in the project to assure a quality level. The validation is carried out by the Validation Centre SPEX. Only those databases whose validation has been positive are accepted by the consortium. Non accepted databases should be updated till a final quality status is achieved.

This section describes the contents of the database and gives a short description of the recording platforms, recording environments and validation procedure

## 2.1. Database contents

The content of the database includes speech utterances to train recognition systems designed for different purposes and applications. The contents of the database can be grouped in:

- Digits, numbers and strings of numbers
- Directory assistance names: Cities, Company names, Forenames, Surnames
- Dates and Times
- Spellings
- Phonetically rich words and sentences
- Application words

Every database comprises recordings of 600 different sessions from at least 300 speakers. 400 sessions consist of the recording of 119 items and the remaining 200 sessions contains 129 items. These 200 sessions includes 10 spontaneous sentences spoken in the car. The sentences are the answer to a specific situation that is explained to the speaker. A total of 22 different situations were described including the management of car accessories, access to databases ... Table 1 shows the contents of each of the 129 items each speaker has to utter. It's indicated if the item is spontaneous or not.

## 2.2. Recording platforms

Two types of recordings compose the database. The first type consist of wideband audio signals recorded directly in the car and the second type is composed by a GSM signal transmitted from the car and recorded simultaneously in a far-end. Two recording platforms were used, a 'mobile' recording platform installed inside the car and a 'fixed' recording platform located at the far-end fixed side of the GSM communications system

The mobile platform records the signals from four high quality audio channels. For this purpose, four microphones were used, a close-talk microphone, and 3 far-talk microphones placed at different locations in the car. The positions for the far-talk microphones are:

A: at the ceiling of the car near the A-pillar
B: at the ceiling of the car in front of the speaker behind the sunvisor
C: at the ceiling of the car over the mid-console (near the rear mirror)

The mobile platform stores the recorded signals as sequences of 16bit, 16 kHz uncompressed and multiplexed. Channels are sequentially multiplexed in short unsigned.

The fixed platform (Fonollosa (2000)) records simultaneously the speech utterances coming from the car through the GSM network (8 kHz sample frequency, A law encoding) . The GSM phone is mounted at the ceiling of the car over the mid-console.

The synchronisation mode between the mobile and fixed platforms is based on use of DTMF tones emitted from the GSM terminal placed in car. A synchronisation

and communication protocol between the two platforms is used to:

- Detecting if PltF is still alive during the recordings (and to repair a hang up);
- Allowing synchronisation of the recordings on the two platforms;
- Allowing the separation of the items in individual files.

The protocols comprise a series of beeps and DTMF-codes transmitted by both platforms to ensure that each recorded item is preceded by a simultaneous beep on all recording channels to allow rapid off-line synchronization of the recordings on both platforms.

Each prompted utterance is stored within a separate file. Each speech file has an accompanying ASCII SAM label file generted both by the fixed and the mobile platforms.

### 2.3. US speechdat-car

The major differences between the American Speechdat-car database and the other databases is that we had to consider the issue of mobile/cell telephony. In order to keep consistency across databases, we decided to collect the data via GSM networks instead of the widespread local standards TDMA and CDMA, although the US GSM operates at 1900 mHz while European GSM operates at 900 or 1800 mHz. The GSM terminal is also different as the one selected for the European collections (the Nokia 5110) that is not available in the USA. We had to choose a "compatible one" (the Nokia 5190). We also selected an American "family" car (Ford Taurus) in which recordings will be conducted.

The textual material for the speaker prompts has also been adapted to US conditions: we decided not to include the word "EURO" as most Americans are not aware of what it refers to and how it is pronounced, we did this also for "Beaujolais" and other typical European words. We also decided not to include the list of major European cities as with the other Speechdat-car databases since it is not expected that many Americans will visit Europe in their cars! So the prompt texts are not a simple adaptation of the British part of the Speechdat-car but constitute the design of an extra language to be incorporated within Speechdat-car collections.

### 2.4. Validation

In order to maintain a high quality level on the databases generated in the project, a validation procedure has been established (Van der Heuvel (2000)) The validation centre SPEX performs some exhaustive checks on the databases:

- Design and completeness
- Formats and Structure
- Annotation quality
- Signal quality

The validation is carried out in two steps. The first step is done in a very early stage of the project and is intended to avoid irremediable errors. In this moment, all the databases have pass this check.

The second step consist on the validation of the complete database and is carried out when the complete database is finished

## 3. Recruiting of speakers

(to be updated by Borge. This is provisionnel)

### 3.1. Recordings of the German database

Both the recording and the annotation of the German SpeechDat-Car database is performed by the Institut für Phonetik und Sprachliche Kommunikation (IPSK) of the University of Munich.

Most of the recording and annotation work in the German SpeechDat-Car is carried out by students who were trained on both tasks. This allows a flexible recording schedule, because most of the time at least one student is available, and an efficient annotation because this work can be parallelized.

Almost all recordings took place between September 1999 and March 2000. Approx. 70% of the recordings were performed in the region of Munich, but experimenters also took the car to other regions for up to one week to achieve the required speaker accent balance. Each recording session took between 30 to 45 minutes, depending on the speaking rate, and the number of item repetitions or GSM connection losses (average 1.5 per session), or other technical problems Draxler (1999-3)].

The recordings of every day are written to a 2 GB Jaz disk in the car and subsequently copied to a file server in the lab. Here, both the signal files and the SAM label files created by the recording software are saved to CD-ROM for backup. All SAM label files are held on the file server to allow a continuous monitoring of the recording progress (gender, age, accent, car environment and item coverage statistics can be computed from the label files). Because of space limitations, the signals of only about 10 recording sessions (~ 2.5 GB) are stored on the server at any one time; they are replaced by new signal files once they have been transcribed.

#### 3.1.1. Problems

Recordings began well before annotation and before a rigorous quality assurance procedure was in place. Consequently, approx. 50 recording sessions contained corrupt data for the close-talk microphone because of a spurious loose contact in the microphone amplifier. When this problem was discovered, experimenters were requested to test the technical quality of the recordings before every single session (instead of only at the beginning of a recording day as previously) and to not start recording if any problem was found (previously, they were allowed to continue recording if they were able to successfully record the test utterance). The high number of lost recordings is due to the fact that a first attempt to repair the loose contact did not fix this problem.

The software for the in-car recordings had severe ergonomical problems (e.g. reproducible error exits) but on the whole it was quite stable. Despite the shock

absorbers on which it was mounted the PC crashed because of strong vibrations, e.g. on gravel or cobblestone roads, or sudden shocks at high speed on highways. Recovery required rebooting the system and a file system check; on two occasions, the recording session had to be aborted because the PC would not reboot.

## 4. Annotation

SpeechDatCar is characterized by having five replicas of each utterance (a close-talk microphone, three in-car microphones, and a microphone connected to GSM). Although the specifications require only annotation of the close-talk channel it is preferable to have annotation tools which allow visualizing and transcribing at the same time of all the given channels. In this way events that occur only in some of the channels can be found. In this section we will describe as an example the annotation procedure used for the German database in detail. Then we will give a short summary for other languages.

### 4.1. Annotation of the German database

The annotation software used is WWWSigTranscribe, an extension of the SpeechDat annotation software developed at IPSK [Draxler (1998); Draxler (1999)]. It is based on the WWW, and it features auditory output of the multi-channel in-car signals as well as the mobile phone channel, and an optional waveform display of all speech signals. To facilitate annotation, editing buttons implement often-needed tasks such as conversion from digits to the appropriate string of digit words, e.g. from ”1 2 3” to ”eins zwei drei”, or to number words, e.g. ”ein hundert dreiundzwanzig”.

For the annotation, the SAM label file created by the in-car recording platform is read in and then renamed, and the prompt text is displayed in the editing field. The annotator listens to the signal output and modifies the prompt text accordingly. Then he or she enters a validation (one of ”ok”, ”bad signal”, ”wrong text” or ”garbage”) and optionally adds a comment. After a lexical check, which ensures that the annotation is syntactically correct, it is saved to a new SAM label file under the name of the original SAM label file.

In principle, the annotation was an auditory annotation of the close talk microphone. However, in order to detect recording problems reliably, annotators were instructed to listen to all four channels for the 10 read sentence items – these items are distributed randomly across recording sessions and thus checking them would reveal technical problems. These 10 items were the first to be annotated so that when a problem was found the annotation session could be aborted early. Furthermore, they were asked to listen to the mobile phone channel for all recordings, and, in case of doubt, check the oscillogram curve.

For the annotation, some of the students who had already worked for the SpeechDat project could be recruited for SpeechDat-Car. Further students were added to the team later in the project. In total, up to 10 part time annotators with between 8 and 19 hours per week were employed. The annotation of a full recording session with the 121 mandatory items plus 10 short spontaneous items plus 9 additional language-specific items in the German data collection takes between 90 minutes for skilled to 150 minutes for inexperienced annotators.

By February 2000, 40.816 signal files of a total of 84.000 signal files had been transcribed (~ 47.7 %). Table 2 shows the number of events found the transcriptions.

| Type | Count | Percent |
|---|---|---|
| signal truncation | 183 | 0.44 % |
| noise markers | 38747 | 46.1 % |
| mispronunciations | 1302 | 3.2% |
| incomprehensible speech | 707 | 1.7% |

Table 2 Events found in the transcriptions of the German Databse

There are noise markers for dial tone, speaker noise (breathing, laughing, coughing), intermediate or stationary noise, or filled pauses. The high number of noise markers can be explained by the prompt beep that precedes every recording which was audible even in many of the close-talk microphone recordings.

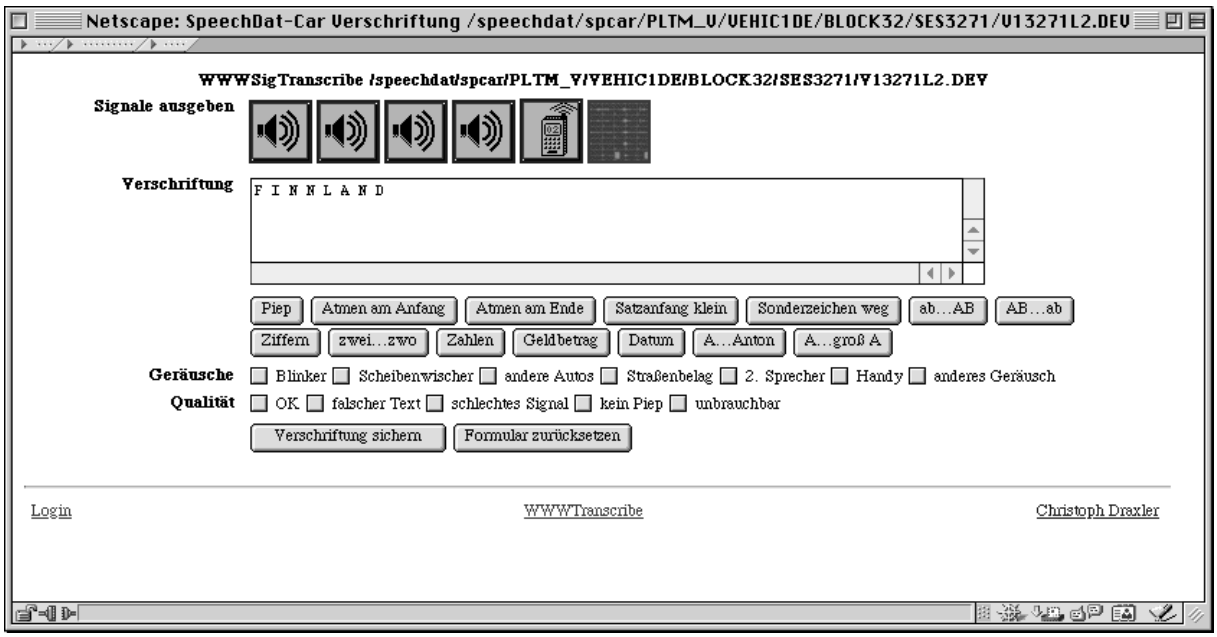
#### 4.1.1. Problems

The annotation of the SpeechDat-Car recordings faced three main problems: adaptation of the SpeechDat annotation guidelines to the SpeechDat-Car requirements, achieving annotation consistency, and software problems.

The original annotation guidelines were designed for quick annotations of telephone speech. The signal quality of the in-car recordings was on the one hand substantially better than telephone speech, but on the other hand much more variable due to the traffic situation, external noises, etc. Hence, the annotation guidelines were updated by the project consortium even after annotations had begun.

Annotation consistency was achieved by a) recruiting a number of annotators that had worked for SpeechDat, and b) asking experienced annotators to function as tutors for new annotators. As such, the veteran annotators would go through one or two recording sessions with the new annotators and monitor their annotations.

The software problems concerned the signal display applet of the WWWSigTranscribe software. Because of a programming error, the applet did not release memory when a signal was no longer displayed and hence quickly filled up the available memory. In many cases a signal display was not needed, and so displaying the waveform was made optional in the final release of the annotation software.




### 4.2. Annotation of the French database

For the French database annotation, the software JavaSgram is used. JavaSgram was developed at IRST with the objective of being flexible for independent annotation of all of the input channels, which can be visualized, zoomed, and unzoomed together in a compact graphic representation (see SpeechCom paper for further information). Currently, 79 sessions have been annotated. On a general basis, there is a fairly low amount of mispronunciations and or truncations (less than 1%) while speaker noises are much more frequent (about 20% to 25 % on the average but it is strongly speaker dependent).

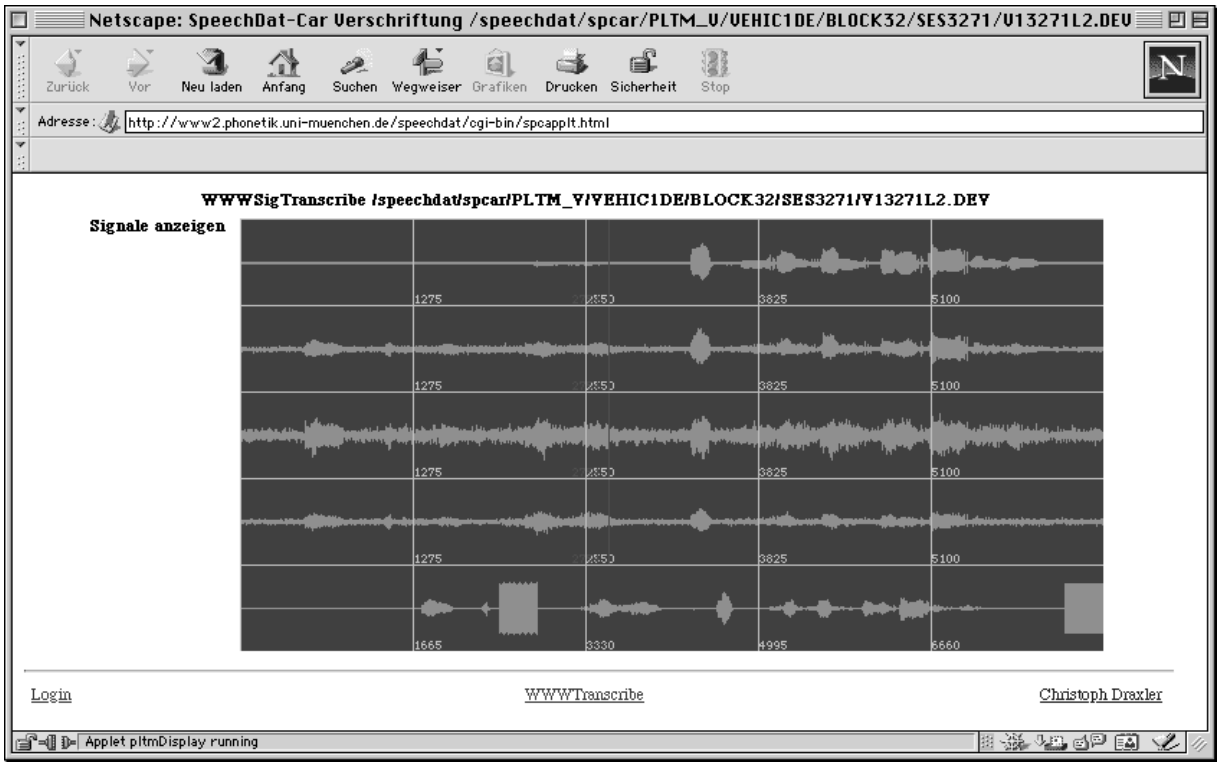
### 4.3. Annotation of the Spanish Database

Only signals from the close talk microphone have been transcribed. Very few errors due to mispronunciation, unintelligible words and truncations have been detected. This is probably due to the direct supervision of recordings in the car and to the fact that the speaker is co-driver and can be concentrated in the recording task. Annotation has been done using the software tool UPCRevBD.v1, developed at UPC (Nogueiras (1998)). The software allows a quick and robust annotation. At the moment of writing this paper, 360 sessions have been annotated. These sessions do not include spontaneous sentences and the annotation time per session is around 30 minutes. Each person is not allowed to annotate more than two hours without rest and no more than four hours per day.

## 5. Distribution

The European Language Resources Association (ELRA) was established as a non-profit association in Luxembourg in February 1995, to provide a European-wide, open platform for the selection and distribution of speech, text and terminology resources to be embedded in language enabled systems, and to promote the use of Language Resources (LRs) within the Human Language Technologies sector. In order to effectively provide such services to research and development groups in academic, commercial and industrial environments, it is necessary for ELRA to address legal, logistic and other practical issues. ELRA has been granted the rights to distribute most of the speech data bases collected within the European funded projects, in particular Speechdat(M), Speechdat-II, Speechdat-East. ELRA will be also trusted with the distribution of the databases being collected within Speechdat-car. As one may imagine, developing such resources is prohibitive, even for large organizations, regardless of the projected market size. Developing a Speechdat-car database is very expensive and time consuming. Linguistic Resources (LR) are universally acknowledged to be critical for the development of robust, broad-coverage, and cost-effective applications for all sectors of HLT, in particular those addressing multilingual issues. It has been decided to make the databases that are produced within the speechdat-car project commercially available via ELRA to third parties that agree to enter into an agreement with ELRA. ELRA will negotiate a distribution agreement with each and every data producer/owner. Third parties will have to enter into only one agreement with ELRA (unless the customer chooses to go to each individual provider and sign as many licenses as languages available playing on several different judicial systems).

The availability rule is clearly stated in the contract that the speechdat-consortium partners signed with the European Commission. It states that, as soon as the speech databases of the different languages are recorded and validated by an independent organization (SPEX, the ELRA validation Unit, see (van den Heuvel (2000)) in these proceedings) they will be available for exploitation to all the other partners of the consortium after they have completed their own databases. All the databases will be distributed to third parties via ELRA no later than 18 months after the official end of the project. In this manner, all the collected speech data will be made available to research institutes and companies all over the world for further exploitation in research and commercial operations.




## 6. Final Remarks

The WWW server of the SpeechDat-Car project is hosted by the IPSK: http://www.speechdat.org/. It contains public and internal deliverables – including all public specifications –, sample recordings, images, videos and country-specific information on the SpeechDat-Car data collections in Europe.

## 7. References


Draxler C (1998); WWWSigTranscribe – A Java Extension of the WWWTranscribe Toolbox, *Proc. of the 1st. Intl. LREC*, Granada, 1998

Draxler C. (1999); WWWSigTranscribe - Annotation via the WWW, *Proc. of MATISSE Workshop,* London.

Draxler C., .Grudszus, R. Euler, S. Bengler K. (1999) First Experiences of the German SpeechDat-Car Database Collection in Mobile Environments, *Proc. of Eurospeech 99*, Budapest,

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.

Fonollosa, J.A.R., Moreno, A. (2000) SpeechDat-Car Fixed Platform *Proc LREC'00*. Athens. Greece

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.*

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.

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.

Nogueiras, A. Moreno, A. (1998) ; NaniBD: A set of Tools for Transcribing and Validating Speech Databases" , *Proc. of the 1st. Intl. LREC,* Granada, 1998

Van den Heuvel , H. Boves,, L. Choukri, K. Goddijn, S. Sanders, E. (2000) SLR Validation: Present State Of Affairs And Prospects *Proc LREC'00*. Athens. Greece

Van den Heuvel, H. Boudy, J., Comeyne, R, Euler, S, Moreno, A, Richard, G. (1999), The SpeechDat-Car multilingual speech databases for in-car applications: some first validation results.