



## SpeechDat – Databases for the Creation of Voice Driven Teleservices

SpeechDat is a CEC-funded initiative (LE2-4001) that addresses the fields of production, standardization, evaluation and dissemination of Spoken Language Resources.

SpeechDat(M) has successfully delivered 8 European language databases of 1000 speakers each. SpeechDat(II) is currently under way. It is a significant extension in that it addresses both short and long term requirements of the telecommunications industry, spoken language technology, and research:

- 20 databases collected in 14 European countries
- recordings for fixed and mobile telephone networks, and speaker verification databases

Each database is validated by SPEX in order to guarantee the quality of the speech material, annotation and documentation. SpeechDat(II) is scheduled to end in February 1998. The databases become available to the public 18 months after completion and final validation. Most SpeechDat(M) databases are already available through ELRA, the European Language Resources Association (<http://www.icp.grenet.fr/ELRA/home.html>)

### SpeechDat contents

3/1 application words/expressions	words and phrases expected in teleservices and applications
6 digit sequences	sheet, telephone and credit card, and PIN-numbers, chain of 10 digits spoken in isolation
2 numbers	isolated digits and natural numbers
1 money amount	local currency expressions
5 date and time expressions	absolute and relative date expressions in numerical and alphanumeric format
3 spellings	real words, artificial letter sequences, proper names
5 directory assistance words	geographical, proper, and company names
2 yes/no responses	spontaneous yes/no responses to questions
4/9 phonetically rich words/sentences	the set of sentences of a recording contains each phoneme at least twice

Signal files are 8 bit 8 KHz headerless files in a-law format. Each signal file has a SAM label file associated to it.

### SpeechDat WWW Site

Visit the SpeechDat WWW server at <http://www.phonetik.uni-muenchen.de/SpeechDat.html> and get the public reports, sound samples from original recordings, partner addresses, and SpeechDat-related publications.

### Contact Address

For further information on SpeechDat, please contact the coordinating Manager at SIEMENS AG, Germany:

Dr. Harald Höge  
SIEMENS AG  
ZFE T SN 5  
Otto-Hahn-Ring 6  
D-81730 Munich  
[Harald.Hoege@mchp.siemens.de](mailto:Harald.Hoege@mchp.siemens.de)

## SpeechDat Databases

Database types are fixed or mobile network recordings, or speaker verification databases. The numbers given show the number of speakers and the number of calls per speaker. The list reflects the status as of April 1997.

Danish	fixed	1000/1	Tele Danmark	completed
	fixed	4000/1	Aalborg University	in progress
Dutch	mobile	250/4	Philips Business Systems	in progress
Dutch (Flemish)	fixed	1000/1	Lernout & Hauspie	completed
English (British)	fixed	1000/1	GEC Marconi	completed
	fixed	4000/1	DPT Ltd.	in progress
	mobile	1000/1	British Telecom	in progress
	verification	120/20	GPT Ltd.	in progress
Finnish	fixed	4000/1	Tampere University of Technology	in progress
French	fixed	1000/1	Philips GmbH	completed
	fixed	5000/1	Matra Communications	in progress
	verification	120/20		in progress
French (Belgian)	fixed	1000/1	Lernout & Hauspie	in progress
French (Luxemburgish)	fixed	500/1	Lernout & Hauspie	in progress
French (Swiss)	fixed	1000/1	IDIAP, Swiss Telecom	completed
	fixed	2000/1		in progress
	verification	20/50		in progress
German	fixed	1000/1	SIEMENS (University of Munich)	completed
	fixed	4000/1		in progress
	mobile	1000/1	Vocalis Ltd. (University of Munich)	in progress
German (Luxemburgish)	fixed	500/1	Lernout & Hauspie	in progress
German (Swiss)	fixed	1000/1	IDIAP, Swiss Telecom	in progress
Greek	fixed	5000/1	Knowledge SA, University of Patras	in progress
Italian	fixed	1000/1	Centro Studi e Laboratori	completed
	mobile	400/1	Telecomunicazione SpA	completed
	fixed	3000/1		in progress
	mobile	250/4		in progress
Norwegian	fixed	1000/1	Telenor AS	in progress
Portuguese	fixed	1000/1	Portugal Telecom SA (INESC)	completed
	fixed	4000/1		in progress
Slovenian	fixed	1000/1	SIEMENS (University of Maribor)	in progress
Spanish	fixed	1000/1	Universitat Politecnica de Catalunya	completed
	fixed	4000/1		in progress
Swedish	fixed	1000/1	Kungl. Tekniska Hogskolan	in progress
	mobile	1000/1		in progress
Swedish (Finnish)	fixed	1000/1	Tampere University of Technology	in progress
Welsh	fixed	2000/1	British Telecom	in progress

## SpeechDat in the WWW

<http://www.phonetik.uni-muenchen.de/SpeechDat.html>