BABEL: A MULTI-LANGUAGE DATABASE

Peter Roach, Simon Arnfield and Elizabeth Hallum

Speech Research Laboratory, Department of Linguistic Science,
University of Reading, UK.

ABSTRACT - A speech database is being constructed by a group of European researchers concentrating on languages of Central and Eastern Europe. The languages covered are Bulgarian, Estonian, Hungarian, Polish and Romanian, and the database is modelled on the Western European EUROM-1 design. The project is co-ordinated by the Speech Research Laboratory at the University of Reading, UK.

INTRODUCTION

Research previously funded by the European Union under the ESPRIT programme has resulted in a well-developed infrastructure and a substantial amount of recorded data from major languages of Western Europe. The SAM project (Fourcin and Dolmazon, 1991) produced protocols for database collection, the design for a speech workstation, a machine-readable phonetic alphabet and a certain amount of specialised software. Ultimately, through building on this work, a multi-language speech database named EUROM-1 was produced (Chan et al, 1995). The European Union's COPERNICUS programme was created to foster collaboration between researchers in existing EU nations and those in countries in Central and Eastern Europe, and the BABEL consortium was formed in 1994 to bid for funding to construct a speech database. The contract (COPERNICUS #1304) for a three-year project was awarded in late 1994 and the project started work in March 1995. It has recently passed its half-way stage.

THE PARTNERS

The consortium comprises partners in Eastern and Western Europe. The Eastern partners carry the main responsibility for planning, recording and labelling the data. The Western partners receive no funding for work of this sort, but act in an advisory capacity on the basis of their previous experience of this type of research. The partners are as follows:

Eastern Europe

Bulgaria: the Phonetics and Speech Technology Group comprising staff from the Institute of Bulgarian Language, Bulgarian Academy of Sciences, the University of Sofia and the Technical University of Sofia. The work is directed by Julia Baltova.
Estonia: the Institute for Cybernetics in the Estonian Academy of Sciences in Tallinn is responsible for work on Estonian under the direction of Einar Meister.
Hungary: the Technical University of Budapest; work is directed by Klara Vicsi.
Poland: two partners are collaborating on the Polish language: the Institute for Fundamental Technical Research of the Polish Academy of Sciences, Warsaw (Ryszard Gubrynowicz), and the Marie Curie Sklodowska University, Lublin (Wiktor Gonet).
Romania: the Technical University of Timisoara; work directed by Marian Boldea.

Western Europe

France: LIMSI, Paris (Lori Lamel); CNRS, Caen (Alain Marchal).
Germany: University of Saarbrücken (Bill Barry); University of Stuttgart (Krzysztof Marasek)
UK: University College London (Adrian Fourcin and John Wells); University of Reading (Peter Roach).

THE LANGUAGES

The languages being worked on in BABEL form a very diverse group, and many of the problems of phonology and symbolisation are unique ones. The following summary is based on Campbell (1991).

Bulgarian belongs to the Eastern branch of Southern Slavonic. There are approximately 8 million speakers of the language living in Bulgaria. It is written with the Cyrillic alphabet (which was originally developed for this particular language), and this creates some problems in the transliteration of example material and in the screen display of reading material. These problems have been overcome by the Bulgarian partners.

Estonian is a member of the Balto-Finnic group of Finno-Ugric. There are approximately 1 million speakers in Estonia, plus a substantial population of expatriate speakers.

Hungarian belongs to the Finno-Ugric branch of Uralic. 12 to 14 million people speak the language in Hungary, and there is a substantial population of Hungarian speakers living in Transylvania.

Polish is the most widely spoken of the BABEL languages, being spoken by approximately 40 million people within Poland and a sizeable number of expatriates. The language is the sole survivor of the Lechitic subgroup of Western Slavonic languages.

Romanian is the only Romance language (i.e. from the Italic branch of the Indo-European family) of the BABEL group. It is spoken by approximately 20 million people within Romania; Moldavian, the language of the Moldovan Republic, is close enough to be considered effectively the same language, and this adds another 2.5 million speakers.

DATABASE RECORDING

The SESAM workstation adopted for this project is PC-based, and incorporates the French-built OROS AU21 board which provides 16-bit analog-to-digital and digital-to-analog conversion, anti-aliasing filtering and high-speed signal processing with the Texas Instruments TMS320C25 processor. Recording is direct to disk with a sampling rate of 20 kHz. The microphone used is the Sony ECM44 (with preamplifier). It was necessary to choose a suitable backup device which would make it possible to send copies of recorded data to different sites by post (since sending such large amounts of data by network file transfer was found to be too slow and unreliable); the device chosen is the Panasonic PD drive, which uses writeable optical disks with 650 mbyte capacity (the drive can also read standard CD-ROM disks). The database is being collected and archived at the co-ordinating site (Reading University) and will be prepared there for ultimate CD-ROM distribution (2 disks per language). Copies of all the material are sent for quality checking in the laboratory of the Warsaw partner, who has this special responsibility in the project.

DATABASE CONTENTS

The project is following the design of EUROM-1 (Chan et al, 1995), though some minor design modifications have been necessitated by specific characteristics of the BABEL languages. The data for each language is recorded in three major sets: (1) the Many-Talker set, (2) the Few-Talker set and the (3) Very-Few-Talker set. The target for (1) is to record 30 women and 30 men reading 100 numbers, 3 connected passages and 5 sentences designed to provide additional instances of phonemes and phoneme contexts that may be inadequately represented in the passages. The target for (2) is 5 women and 5 men (selected from the Many-Talker set speakers) reading sets of syllables representing the syllable structure of the language, 5 sets of 100 numbers, 15 passages and 25 "filler" sentences. In (3), one woman and one man read syllabic material embedded in 5 different context phrases and 5 x context words. The final database documentation will give the full specification of the material.

ANNOTATION

It is of the greatest importance that the acoustic data should be accompanied by the fullest possible annotation. It is, however, not practical to attempt to label the entire database at phoneme level, and it has therefore been decided to concentrate on labelling phonemically the Connected Speech passages of the Many-Talker set. This will result in approximately 1.5 hours of phonemically labelled speech per language. All the other material is effectively end-point marked as a result of the recording technique. This uses the EUROPEC package, another product of the SAM project. This software interacts with the OROS AU21 board to control the recording process. The text to be recorded is typed in and stored, and presented visually to the speaker at pre-set rates; the timing of the material is recorded along with the acoustic signal so that individual data items are identifiable on the time course of the file in which they occur.

Segmental labelling presents many problems. Some are theoretical and some are practical. Dividing speech into temporally non-overlapping segments is itself controversial (Roach et al, 1990). There is the issue of the meaning of 'phonemic' in this context - what is phonemic to a theoretical phonologist is very different from the phonemic level as seen by laboratory phoneticians (Barry and Grice, 1991). There are advantages in labelling in a more 'phonetic' or 'allophonic' mode which labels contextual variants of sounds explicitly, but there is a cost in transcriber time. Calculations for planning labelling time have been based on a notional figure of 100 times real time (i.e. 1 minute of speech takes 100 minutes to label). Labelling requires a machine-readable phonetic alphabet, and the SAMPA conventions provide a base for the choice of symbols for our database. These conventions are largely the work of J.C. Wells (see for example Wells, 1995), and may be viewed via World Wide Web at http://www.phon.ucl.ac.uk/home/sampa/home. The symbols chosen for each language are set out below; where there are major allophonic differences, diacritics may be added to indicate such features as voicelessness, affrication or vocalization.

Bulgarian
vowels: i , e , a , @ , O , u
consonants: p , b , t , d , k , g , ts , dz , tS , dZ , f , v , s , z , S , Z , x , m , n , l , r , j
p' , b' , t' , d' , k' , g' , ts' , dz' , f' , v' , s' , z' , x' , m' , n' ,l' , r'

Estonian
vowels: i , ii , e , ee , { , {{ , y , yy , 2 , 22 , u , uu , o , oo , 7 , 77 , A , AA
consonants: p , pp , t , tt , k , kk , t' , t't , f , ff , v , vv , s , ss , S , SS , h , hh , s' , s's , m ,
mm , n , nn , n' , n'n , l , ll , l' , l'l , r , rr , j , jj

Hungarian
vowels: i , i: , E , e: , O , A: , o , o: , 2 , 2: , u , u: , y , y:
consonants: p , b , t , d , t' , d' , k , g , ts , dz , tS , dZ , f , v , s , z , S , Z , m , n , J , r , l , j ,
h , x

Polish
vowels: i , I , e , a , o , u , e~ , o~
consonants: p , b , t , d , k , g , f , v , s , z , S , Z , s' , z' , x , ts , dz , tS , dZ , ts' , dz' , m , n ,
n' , N , l , r , w , j

Romanian
vowels: i , i_0 , e , a , @ , o , u , 1
consonants: p , b , t , d , k , g , ts , tS , dZ , f , v , s , z , S , Z , h , m , n , l , r

COMPLETION OF DATABASE

Completion is scheduled for early 1998. It is intended that the database will be distributed by ELRA, the European Language Resource Association. Anyone wishing for information about progress on the project is advised to check the BABEL World Wide Web pages, which are regularly updated: http://midwich.reading.ac.uk/research/speechlab/babel

CONCLUSION

Speech technology is widely predicted to be the major development in human-computer interaction in the next century. Development of suitable systems for recognition and synthesis requires large amounts of very high quality data for training and testing. The developing economies of Eastern Europe will certainly be among those anxious to be involved in such a technological revolution, and the BABEL project represents an important enabling initiative.

REFERENCES

Barry, W. and Grice, M. (1991) 'Problems of transcription and labelling in the specification of segmental and prosodic structure', *Proceedings of the XII International Congress of Phonetic Sciences,* Aix-en-Provence, Vol.5, pp.66-69.

Campbell, GL. (1991) *Compendium of the World's Languages,* (Routledge).

Chan, D. and others (1995) 'EUROM - a spoken language resource for the EU', *Proceedings of Eurospeech,* Madrid, Vol.1, pp. 867-870

Fourcin, A.J. and Dolmazon, J-M. (1991) 'Speech knowledge, standards and assessment', *Proceedings of the XII International Congress of Phonetic Sciences,* Aix-en-Provence, Vol.5, pp.430-433

Roach, P., Roach, H., Dew, A. and Rowlands, P. (1990) 'Phonetic analysis and the automatic segmentation and labelling of speech sounds', *Journal of the International Phonetic Association,* vol.20.1, pp.15-21.

Wells, J.C. (1995) 'Computer-coding the IPA: a proposed extension of SAMPA', *Speech, Hearing and Language,* Work in Progress, Department of Phonetics and Linguistics, University College London.

The German SpeechDat Telephone Speech Corpus
Overview and Experiences

Christoph Draxler

Department of Phonetics and Speech Communication
University of Munich

ABSTRACT
SpeechDat is a European project to collect ISDN quality telephone speech for all major European languages. In the first phase of the project, 1000 speakers were recorded in eight languages, including German. The paper presents the experiences made during the data collection for German, and outlines the specifications for the second phase of the project.

INTRODUCTION

The SpeechDat[1] project is a joint industrial and academic effort to collect Polyphone-like ISDN quality telephone speech corpora for major European languages. These corpora provide a common basis for the development of speech processing applications for the European languages, and for phonetic, linguistic, and telecommunications research.

The project is divided into two phases, SpeechDat(M) from September 94 until February 96, and SpeechDat II from March 96 until March 98. SpeechDat (M) served primarily as a case study for the definition of common standards, the technical feasibility of the data collection, and the creation of a data validation and distribution infrastructure. These goals have been achieved successfully:

- the ELRA (European Language Resource Association) has been established and funding is guaranteed by the CEU for three years, and
- all SpeechDat(M) corpora have been validated and pressed on CD-ROM.

The main objective of SpeechDat II is to extend the data collection to larger speaker populations and to include additional languages.

Language	SpeechDat (M)	SpeechDat II	Language	SpeechDat (M)	SpeechDat II
British English	1000	4000	Welsh		2000
Danish	1000	4000	Belgian French		1000
German	1000	4000	Finnish Swedish		1000
Italian	1000	3000	Flemish		1000
Portuguese	1000	4000	Norwegian		1000
Spanish	1000	4000	Russian		1000
French		5000	Slovenian		1000
Greek		5000	Swiss German		1000
Swedish		5000	Luxemburgish French		500
Finnish		4000	Luxemburgish German		500
Swiss French		2000			

Table 1. Speechdat Languages and Speaker Populations

A similar data collection with 5000 callers has been carried out for Dutch prior to the SpeechDat project (den Os et al.,1995). 1000 callers were recorded for Swiss French at the same time when SpeechDat(M) began; this corpus was later included in SpeechDat(M).

Structure of the paper

The remainder of the paper is structured as follows: section 2 contains the SpeechDat corpus specifications, section 3 outlines the technical setup. Section 4 describes the speaker recruitment and contains statistical information on the speaker population. Section 5 presents the orthographic transliteration. Section 7 discusses the SpeechDat(M) validation, and section 8 summarizes the experiences and gives an outlook on SpeechDat II.

SPEECHDAT DATABASE SPECIFICATIONS

The German SpeechDat(M) corpus consists of 1000 calls. Each call consists of 44 recorded items, 39 of which were mandatory (for all languages) and 5 were optional. In SpeechDat II, the emphasis has shifted towards a better phoneme and diphoneme coverage, and towards real-world speech data such as geographical and company names; furthermore, the application words and phrases have been revised to match the requirements of telephony applications (Winski, 1996).

SpeechDat		Item Type	Specification
M	II		
1	2	isolated digit items	isolated digit, sequence of 10 isolated digits
3	4	digit/number strings	prompt sheet number, telephone number, credit card number, PIN code
3	1	natural number	
2	1	money amount	currency amount, mixed size and units
3	2	yes/no questions	spontaneous replies
3	3	dates	spontaneous date, e.g. birthdate, prompted text form, relative and general date form
2	2	times	spontaneous and prompted time of day, mixed analogue and digital format
6	3	application keywords/keyphrases	
3	1	word spotting phrase using embedded application words	
1	4	directory assistance names	city of birth/growing up (spontaneous), most frequent cities, most frequent companies or agencies, forename (spontaneous)
	1	proper name	set of 150 SDB full names
3	3	spellings	spelling (spontaneous), directory city name, real/ artificial word
0	4	isolated words	
9	9	phonetically rich sentences	
5	11	partner specific material	speaker gender, fuzzy question, birthdate, speaker region, today's date, form task, spontaneous speech, good-bye phrase
44	51	TOTAL	

Table 2. SpeechDat Item List

The exact vocabulary to be recorded was specified separately for each language. In SpeechDat(M), the most restrictive constraint for the corpora was that every prompt sheet must contain each phoneme of the language at least twice (except for very rare phonemes). In most cases this was achieved by selecting the phonetically rich sentences from a large text corpus, e.g. newspaper text.

As a general guideline, the duration of an interview should not exceed 10 minutes. For German, the total duration of an interview was approx. 8 minutes. In a complete interview, approx. 4 minutes of speech were recorded.