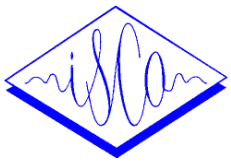

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

# BULGARIAN SPEECH DATABASE: A PILOT STUDY

A.Misheva (Bulgarian Academy of Sciences), S.Dimitrova (University of Sofia), V.Filipov (Universities of Leeds and Sofia), E.Grigoreva (University of Sofia), M.Nikov (University of Sofia), P.Roach and S.Arnfield (University of Reading)

## ABSTRACT

The paper describes the construction of a database of Bulgarian speech which follows the protocols which were established by the ESPRIT SAM project and have become an accepted European standard for such work. A number of different transcription systems were attempted. Future work on other languages of Central and Eastern Europe will build on the foundations laid by the work reported here.

## 1. INTRODUCTION

There is a growing quantity of recorded speech material that has been prepared in a form suitable for use as a computer database in speech technology research. This work has been principally concentrated on widely-used languages such as English, French, German and Japanese, but for many other languages there has been so far little or no progress. The authors are members of a new COPERNICUS project (Project 1304 "BABEL: a Multi-Language Database") which will create a database of spoken Bulgarian, Estonian, Hungarian, Polish and Romanian. This work is only in its early stages, but a preliminary study of Bulgarian, whose design is intended to form a basis for the main project, was initiated some time ago by the Phonetics and Speech Technology Group in Sofia, and the work of that study is described in this paper.

## 2. DATABASE STANDARDS

Our intention was to follow as fully as possible the standards laid down by the SAM project (ESPRIT project no. 2589) [1]. This project devised a set of protocols suitable for database work on any European language, and includes the specification of a speech workstation (SAM-Station), a machine-readable phonetic / phonemic alphabet (SAMPA), software for presentation of material to speakers and file creation (EUROPEC) and conventions for file formats, recording conditions, safety backup of recordings and so on.

## 3. RECORDING

It was decided that the protocols for speech recordings established by the SAM project should be used throughout this study. Since no facilities existed in Sofia suitable for making recordings of the required standard it was decided to record two native Bulgarian speakers in the UK, one female (MK) and one male (PK); these speakers were in their mid-40's. The recordings were carried out in the recording studio of the Department of Linguistics & Phonetics at the University of Leeds. A PC 386 computer was used, which was originally equipped with an OROS AU-21 single-channel signal input/output and processing board. It was decided that for optimum conformity with SAM protocols a two-channel recording should be made, with a laryngograph signal recorded on the second channel. The laryngograph channel is optional in SAM protocols, but was included for possible later use in pitch-synchronous analysis of the acoustic data. For this purpose the Department of Phonetics and Linguistics at University College, London, kindly lent a dual-channel AU22 board for the duration of the recording process.
During the recordings the speaker sat in a purpose-built room designed to minimise reverberation, with laryngograph electrodes attached to the neck, and with

an AKG condenser microphone positioned 10 cm from the lips. The recording computer was positioned in an adjacent room which was acoustically isolated from the recording room. The signal levels were monitored by the person operating the computer, and were adjusted for each recording session so that the peak was always at least 10dB below the recommended maximum signal amplitude. The recording process was controlled by the Europec software package which was produced for this purpose by the SAM project. This allows the user to input a list of stimulus items together with information about the recording conditions, speaker identity and data type. The recording part of the program then gives the speaker a visual prompt, records for a set time, files the newly-acquired data, pauses for a pre-set time and then moves on to the next item. It is, of course, not possible to have the computer in the room with the speaker because of the noise it generates; after experiments with a second computer monitor, it was decided instead to use a TV monitor relay, with a small television showing the speaker a close-up picture of the relevant part of the computer display, as seen by a TV camera. One problem that had to be overcome was that it was not possible to display the Cyrillic characters in which Bulgarian is written, and the speakers were not phonetically trained. It was therefore necessary to convert the Cyrillic writing into a Romanised equivalent and then train the speakers to read this without hesitating. The data acquisition system used works with 16-bit samples, and 20kHz was chosen as the sampling rate. Compared with many speech databases this may seem an extravagantly high rate (16 kHz being more common), but the recordings made by SAM used this sampling frequency and it was important to achieve the highest quality of recording that was practical. The OROS board automatically applies appropriate anti-aliasing filtering. Recording "takes" were limited to about 30 sec to avoid the risk of overrunning the computer's memory capacity (4 mbyte); each "take" thus generated two files (one for each channel), each of about 1 mbyte.

## 4. TRANSCRIPTION

It was decided that the computer analysis of the recordings should be carried out in Sofia by the Bulgarian phoneticians in the group. Ideally, the data would have been kept in digital form throughout the study, but for practical reasons it was necessary to transport the recordings to Sofia in audio form. The files were played through the OROS board's d-to-a convertor on to a professional-quality stereo tape recorder, and re-digitised in Sofia on the Phonetics Laboratory's PDP-11/73 computer using a 12-bit a-to-d converter. An interactive signal analysis and editing program was used to identify segments and segment boundaries, and the symbols for these were stored in transcription files using SAM format. The transcription files were then returned to Leeds to be matched to the corresponding signal files on the original PC speech workstation used for making the recordings. Some adjustments were necessary for two reasons: firstly, the onset of the signal in the files was not perfectly synchronised between the Leeds and Sofia versions, and therefore a timing offset had to be calculated and applied, and secondly a very slight tape speed difference between the two sites meant that a further small correction factor (around .75%) had to be applied. After these corrections had been carried out, detailed checks were made on the files and the alignment appeared to be as good as could be achieved in normal computer-based transcription practice.

There is no single accepted set of phonetic or phonemic transcription systems for speech database work. At the 1989 Kiel Congress of the International Phonetic Association it was decided to have a code number identifying every possible I.P.A. symbol and diacritic. Instead of using the usual codes for computer symbols (ASCII), which offer rather limited possibilities for designing new symbols, they devised a completely new coding. The latest version of this is presented in **[2]**. The problems and principles of computer-based transcription are explained briefly in **[3]**. A different activity has been the

development of *machine-readable* phonetic alphabets that can record all necessary phonetic detail using characters available on the keyboard of an ordinary computer. Several such alphabets have been devised, but the most comprehensive is the SAMPA system devised by John Wells for the ESPRIT SAM project [4], which is being progressively extended to cover the phonetic features of any language. Since nothing of this sort has previously been attempted for Bulgarian, it was decided to transcribe the data with a variety of symbol sets and then subsequently evaluate how well suited each one was.

## 5. DATABASE CONTENTS

Given the goals of the pilot project the following criteria were employed in the compilation of the database:

> (i) the database should be both quantitatively limited and yet representative of the segmental and suprasegmental system of Contemporary Standard Bulgarian (CSB);
> (ii) it had to comprise data similar to those used in the SAM project.

The first section comprises the cardinal numbers from 1 to 10. A second section contains all the consonant phonemes of Bulgarian in intervocalic position with a fixed initial and final vowel /a/, the stress being on the second syllable.

List No.3 consists of the stressed and unstressed vowels of CSB included in nonsense words of CVCV structure, C being fixed (/p/) and the stress occurring consecutively on the first and the second syllable. This is done in view of the vowel reduction in CSB.

List No.4 consists of six sentences illustrating different communicative functions: two statements, three different types of question, and a command. The questions have almost identical segmental structure, thus more clear-cut intonation pattern differences are expected. The segmental makeup of the six sentences covers all the vowels of CSB as well as 16 of the 22 non-palatalised consonants.

With a view to enhancing the diagnostic power of the study two additional lists were included. Thus list No.5 contains 41 words of CSB of CVC structure representative of different consonants occurring in initial and final position. The vowel is /a/, except in five of the words where a front vowel was chosen, thus incorporating the palatalised allophones of /k, g, x, l/.

List No.6 consists of 11 CVCV words where the consonant in both positions is /p/, the different vowels occurring both in stressed and unstressed position.

The study has thus resulted in a small database of Bulgarian speech which is, of course, entirely in digital form. The authors would be happy to allow other researchers access to this material. Its contents are as follows:

Speech signal files; Laryngograph signal files; File definition files; Phonemic transcription files; "SCRIBE" transcription files; Acoustic segment-based transcription files; Prosodic transcription files.

## 6. ANALYSIS OF RECORDINGS

A good test of the accuracy of labelling of such material is the success with which a recognition system can be trained on it; experiments are in progress to train Hidden Markov models on the segments in the database.

## REFERENCES

[1] Fourcin, A.J. and Dolmazon, J-M 'Speech knowledge, standards and assessment', in *Proceedings of the XII ICPhS*, Vol.5, pp.430-3, University of Aix-en-Provence, 1991.

[2] Esling, J. and Gaylord, H. 'Computer codes for phonetic symbols', *Journal of the I.P.A*, vol.23 no.2, 1991, pp. 83-97.

[3] Grice, M. and Barry, W. 'Problems of transcription and labelling in the specification of segmental and prosodic structure' in *Proceedings of the XII International Congress of Phonetic Sciences*, Aix-en-Provence, vol. 5, 1991, pp. 66-9.

[4] Wells, J.C. 'Computer-coding the IPA: a proposed extension of SAMPA', *Speech, Hearing and Language,* University College London, 1994, pp. 271-289.