A brief history of the Romanian language phonetic databases.
(A note on previous researches related to databases for the Romanian language)

(Draft article. By Horia-Nicolai Teodorescu)

The researches to produce databases have in Romania, as in any country, a longer history, starting much before the advent of computers. The first phonetic databases have been produced in Europe in the 19th century, when the recording technology has appeared. In Romania, most early databases we are aware of have been related to dialectology and folklore research. Probably the oldest such database is preserved at the Romanian Academy, on discs (plates). A specialized institution of the Romanian Academy, Arhiva de Folclor a Academiei Române (The Archive for Folklore of the Romanian Academy), from Cluj, (see www.academiaromana.ro/academia2002/acadrom/pag_inst.htm) is entitled to host such recordings, among other institutions. Most probably, there are speech databases at the Institutul de Fonetică şi Dialectologie (Institute for Phonetics and Dialectology) "Al. Rosetti", from Bucharest. Some other phonetic databases regarding dialects are preserved in the major universities, most of them on magnetic tapes.

Several groups of authors have reported the establishing of phonetic Romanian language databases. We would like to acknowledge these researches and these contributions, yet deploring that the data on the databases is scarce and that the databases are not public, at our best knowledge. Also, because these databases have been produced with specific applications in mind and because they have not been intended for public distribution, their documentation is probably tailored only for the use of the authors of the respective databases. Some of these researches are listed subsequently. We are not aware of any public free database related to these researches.

Felicia Serban, Dana Bucerzan, Luciana Peev, and Lidia Bibolar, have reported on a "Database of the Romanian Language Phonetics and Phonology", http://www.racai.ro/books/awde/serban.html.

Another reporting is by Corneliu Burileanu & al.: Text-to-Speech Synthesis for Romanian Language. http://www.racai.ro/books/awde/burileanu8.html. According to the authors:

> "The acoustic data-base contains up to the present more than 4000 syllables digitally stored. They were introduced in a single file, but with many index files, so that the searching is very fast. The syllable is found in two steps: first, the index file referring to the syllable is found, then the syllable position in the data file is found, based on information provided by the index file. This searching algorithm was conceived using the C++ programming language, with eight index files. The data-base structure was organized in terms of phoneme number composition."

A sound database is reported in Attila Ferencz & al., A Text-To-Speech System for the Romanian Language. http://www.racai.ro/books/awde/ferencz1.html. The authors state that:

> "The sound database of our system consists of 900 diphones."

In the paper "Speech Technology Research at Computer Science Department, "Politehnica" University of Timisoara", the author, Marian Boldea, briefly describes a small database for words representing the first 10 digits:

> "… the first linguistic resource we collected was a small speech database … consisting of 300 signal files containing isolated utterances in Romanian of 0 to 9 digits, repeated three times by 100 speakers (50 men and 50 women), plus a file containing for each speaker some personal data about aspects that had or could have had an influence on their voice quality: sex, height, weight, age, mother tongue, smoking habits, etc. Organised in two training and test sets consisting of 68 and 32 speakers respectively, and partly hand-labelled at the word level, this database has been used for speech recognition studies, but its use is by no means limited to this, and we can make it available to other interested laboratories."

Finally, the first author of this site / database has used to build a tape-recorded sound database for vowels and words representing the first 10 digits. The recordings have been made during the years 1980-1990, for specific purposes (determining the vowel triangle, medical applications, speech recognition). Some results have been reported in several papers and in the booklet "Man-Machine Communication", Editura Tehnica, Bucharest (authors H.N. Teodorescu, L. Buchholtzer, C. Posa). Also, a small computer-based speech archive has been produced for educational purposes by the same author during the years 1995-2001. Unfortunately, as with many other speech databases, the lack of annotation and fluctuations in the technical conditions for recording made parts of those databases unsuitable for inclusion in the present Internet-based archive.

There is a speech database for the Romanian language on the Internet, namely the BABEL Romanian database, made commercially available by the ELDA Consortium (http://www.elda.org/rubrique1.html). According to the catalog of ELDA, "The BABEL Database is a speech database that was produced by a research consortium funded by the European Union under the COPERNICUS program (COPERNICUS Project 1304). The Romanian database consists of the basic "common" set which contains the Many Talker Set (50 males, 50 females), the Few Talker Set (5 males, 5 females), and the Very Few Talker Set (1 male, 1 female)." (http://www.elda.org/article.php?id_article=18#S0170). The price tag varies, depending on the customer and the use (research or commercial applications) from 300 to 6000 EURO. Unfortunately, no detailed information is offered about protocols (recording conditions and qualities, subjects, corpus etc.). Moreover, the authors provide no information about the database (recording conditions and qualities, subjects – speakers' profiles, corpus etc.). What we can determine is from the Quick Quality Check Report. Among others, "No information on speaker accents provided.", "No information on environment distributions provided.", "No lexicon provided." (See http://www.elda.org/catalogue/en/speech/qqc/s0170qqc.pdf). Therefore, we are unable to compare our database with the Babel one.

The speech archive we propose is a free database, aimed to encourage learning and research of the Romanian language. Other distinctive features of our database are detailed on this site, in several articles.

Resources

European Language Resources Association Home http://www.elra.info/
ELDA - Evaluations and Language resources Distribution Agency web site: http://www.elda.org/article158.html