A Corpus of the Sounds in the Romanian Spoken Language for Language-Related Education

Horia-Nicolai Teodorescu*,**,***
Diana Trandabăț*, ***
Monica Feraru**
Marius Zbancioc*,**
Ramona Luca*
* Romanian Academy, Institute for Computer Science, Iaşi, Romania
** CERFS Center for Research, Technical University “Gh. Asachi”, Iaşi, Romania
*** Faculty of Computer Science, University “Al.I. Cuza”, Iaşi, Romania

Overview

- Introduction
- Description of the corpus
- Methodology of sound acquisition
- The speakers and the voices
- Corpus annotation
- Analysis instruments
- Educational outreach
- Conclusions

Introduction

- The aim is to create an educational and scientific free database (corpus) for the spoken Romanian language
- The corpus is extensively documented, in order to allow both scientific research and informed education based on it.

Introduction

This site can be used

- for educational purposes
- for the analysis of sounds
- for the analysis of specificities of the Romanian language pronunciation compared to other languages
- for research purposes (including medical researches)

Introduction

- The corpus provides Romanian and foreign teachers, linguists, students and researchers a database of medium to high quality, for the study and analysis of the Romanian language sounds
- We provide a tool for international knowledge, learning, and recognition of the Romanian language phonation and to initiate a national Web archive for the Romanian language

Description of the corpus

- The database includes both professional voices ("perfect" pronunciations), as well as non-professional voices
- The database includes a section with files containing syllables and words pronounced in various contexts
- The database includes a section with files of sounds, syllables and words pronounced by persons with various pathologies

Description of the corpus

- We aim to include sections with dialectal pronunciations
- A further goal is to achieve multimedia Dialectal Romanian Language Atlases, with databases of specific pronunciations

Description of the corpus

- We aim to produce a vast systematical study of the currently spoken Romanian language.
- This includes
  - a statistical vowel triangle
  - statistical characteristics of the spectra
  - regional statistical characteristics (dialectal) etc

Description of the corpus

- The database is aimed to serve as a basis for building concatenative voice synthesizers
- The database may be helpful in improving voice recognition systems based on acoustical features

Methodology of sound acquisition

- The acquisition of the voice signals has been performed following a well-determined protocol regarding the environmental noise level and the acquisition hardware
- The recordings have been made with a PC with a motherboard type MB FOXCONN 760 GXK8MC-S, with an incorporated soundboard Sound MAX Digital Audio manufactured by Analog Devices

Methodology of sound acquisition

- The recordings were realized in a speech laboratory, with reduced noise, but not in a complete phonically isolated environment
- The microphone has an essential role in the quality of the recordings
- The position is below the mouth

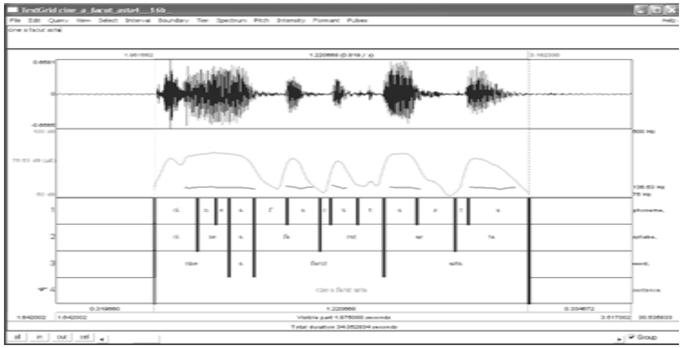
The speakers and the voices

- The voices included in the corpus belong to young subjects with higher education, all 20 to 30 year old, most of them not professional speakers
- All speaker come from the Moldova region

POLAND
SLOVAKIA
UKRAINE
CARPATHIAN MOUNTAINS
HUNG.
MOL.
Oradea
Cluj-Napoca
TRANSYLVANIA
Iasi
Bacau
Târgu-Mureş
Arad
Timişoara
Sibiu
Moldoveanu
Braşov
Galati
Braila
TRANSYLVANIAN ALPS
Ploieşti
Piteşti
Tulcea
BUCHAREST
WALACHIA
SER. AND MONT.
Craiova
Constanța
Danube
Giurgiu
Mangalia
BULGARIA
Black Sea
0 50 100 km
0 50 100 mi
20
24
28
48
44

Corpus annotation

- Manual annotation at different levels: *Phonetic, Linguistic, Emotional, Medical, etc.*



Corpus annotation

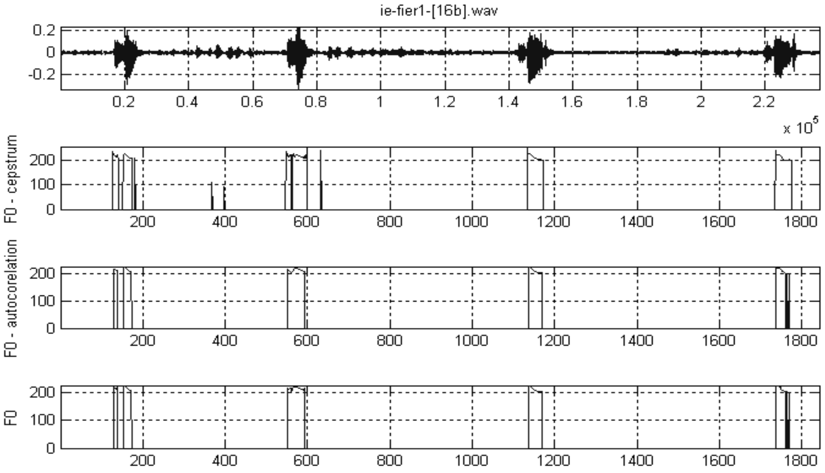
- The information on the phonological level will be completed with prosodic information (tone, intensity, increase units, etc)
- One of our goals is to distinguish the prosodic features which make the difference between human and synthetic speech

Analysis Instruments

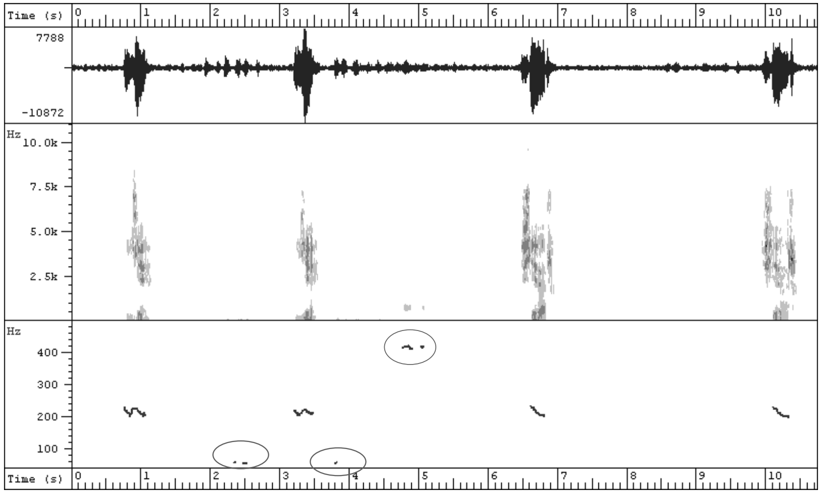
The tools already available are:

- the instant spectrum
- the medium spectrum
- the pitch detection using:
  - the cepstral method
  - the autocorrelation
  - a combination between the two methods.

Analysis Instruments



Analysis Instruments



Educational outreach

- The corpus is used in teaching and laboratory activities in the class “Speech Technology” given for the master degree in “Computational Linguistics”
- The corpus is complemented by a volume (in Romanian) on speech signal analysis

Educational outreach

We hope that it will be used

- in all the universities in Romania where foreign students learn the Romanian language
- in other academic media
- an online tool by foreign students and teachers

Conclusions

- We have created incorporates a quite large corpus of phonemes, words and phrases spelled by a significantly large number of subjects within various contexts
- The corpus is incremental and new recordings with the corresponding annotations and documentations will be continuously added

Conclusions

A large range of tools are provided:

- an extensive documentation on the subject condition
- peculiarities
- spelling conditions
- recording conditions

The Internet allows the international community to use and to compare language analysis and research

Thank you !