Title

The Romanian ASR System developed at “Speech & Dialogue” (“SpeeD”) Research Laboratory - University POLITEHNICA of Bucharest. Under-resourced languages challenges for an ASR system.

Topics

- acoustic modeling;
- text pre-processing;
- language modeling;
- pronunciation modeling.

Abstract

State-of-the-art Automatic Speech Recognition (ASR) systems for high-resourced languages use hundreds or even thousands of hours of *manually transcribed speech data* for training the acoustic model and corpora with billions of words to create the language model. This is a critical issue in the development of a new ASR system, because the acquisition of such data is expensive and requires a lot of time. Under-resourced languages are characterized by lack of text corpora and annotated speech data, phonetic dictionaries, tools and language expertise.

This presentation represents an overview of the Speech and Dialogue research group contributions to *ASR adaptation for under-resourced languages* (more specifically for the Romanian language). Most of our contributions are *language independent* and can be applied to any other under-resourced language, but some of them are *specific to Romanian*. Among the Romanian-specific contributions, we discuss (i) the acquisition of the largest text corpus to be used for language modeling and (ii) the development of the first large-vocabulary ASR system for Romanian and (iii) diacritics restoration for online texts, as a mandatory text pre-processing operation. Our language-independent contributions concern (i) ASR domain-adaptation using statistical machine translation to create domain-specific language models, (ii) automatic grapheme-to-phoneme conversion for pronunciation modeling and (iii) semi-supervised and unsupervised acoustic model training.