Contribuţii la un sistem de recunoaştere de vorbire continuă, cu vocabular extins, independent de vorbitor pentru limba română

- Abstactul Tezei de doctorat -

Această teză de doctorat prezintă contribuţiile aduse de autor la construcţia unui sistem de recunoaştere a vorbirii continue cu vocabular extins (RVC-VE), independent de vorbitor, pentru limba română. În principal din cauza lipsei de resurse acustice, fonetice şi lingvistice un astfel de sistem nu a fost dezvoltat încă pentru limba română. Deşi pentru limbi precum engleza sau franceza sistemele de RVC-VE sunt utilizate pe scară largă de mai bine de un deceniu, pentru alte limbi mai slab dotate (din punctul de vedere al resurselor mai sus menţionate) sistemele de recunoaştere a vorbirii au performanţe acceptabile numai în condiţii de vocabular redus, gramatici specifice unei sarcini simple, etc.

Una dintre principalele contribuţii ale acestei teze a fost achiziţia şi preprocesarea resurselor necesare pentru dezvoltarea unui sistem de RVC-VE. Au fost achiziţionate mai multe baze de date de vorbire care reprezintă în acest moment cel mai mare material vorbit şi etichetat în limba română disponibil pentru cercetare. De asemenea au fost colectate mai multe corpusuri de text de dimensiuni mari ce cuprind ştiri publicate pe ziarele online româneşti. În vederea preprocesării materialelor text colectate de pe Internet au fost proiectate şi implementate un utilitar de curăţare a textului şi un sistem de restaurare a diacriticelor. Dicţionarul fonetic iniţial nu a fost suficient pentru un sistem cu vocabular extins şi a fost suplinit cu ajutorului unui sistem automat de fonetizare.

Utilizând resursele astfel obţinute, au fost dezvoltate un model acustic adaptat vorbirii continue în limba română, un model de limbă şi un model fonetic pentru limba română. Acestea sunt componentele de bază ale unui sistem de recunoaştere a vorbirii continue. Ele au fost integrate şi folosite pentru a dezvolta primul sistem de RVC-VE pentru limba română. Independenţa de vorbitor a sistemului a fost şi ea evaluată şi s-a ajuns la concluzia ca sistemul recunoaşte foarte bine vorbirea rostită majoritatea vorbitorilor. Există însă şi voci mai speciale pentru care performanţa de recunoaştere nu este suficient de bună. În concluzie, deşi mai sunt o serie de optimizări ce trebuie abordate, sistemul prezentat în această teză poate fi considerat primul sistem de RVC-VE independent de vorbitor creat pentru limba română.

Towards a speaker-independent, large-vocabulary continuous speech recognition system for Romanian

- PhD Thesis abstract -

This PhD thesis presents the author's contributions towards developing a speaker-independent, large-vocabulary, continuous speech recognition (LV-CSR) system for Romanian. Mainly due to the lack of acoustic, phonetic and linguistic resources, such a system has not been developed yet for the Romanian language. While for languages like English or French LV-CSR systems are widely used for more than a decade, for many other under-resourced languages (in terms of the above resources) speech recognition systems perform acceptable only in some specific cases: small vocabulary, simple task-grammars, etc.

One of the main contributions of this thesis was the acquisition and preprocessing of the resources needed for the development of the LV-CSR system. We have collected several speech databases which currently form the largest spoken and labeled Romanian material available for research. We also collected several large corpora of text that contain news published in some Romanian online newspapers. In order to process the text collected from the Internet we have designed and implemented a text cleaning tool and a diacritics restoration system for Romanian. The initial phonetic dictionary was not enough for a CSR system with large vocabulary. Consequently, an automatic fonetization system was developed to help the phonetization of words in new vocabularies.

Using the collected resources, we have been developed an acoustic model adapted to continuous speech in Romanian, a language model and phonetic model for Romanian. These are the basic components of a continuous speech recognition system. They were integrated and used to develop first LV-CSR system for Romanian. The speaker-independence of this system was also evaluated and we have concluded that the system recognizes very well the speech uttered by most of the speakers. However, there are soem special voices for which the recognition performance is still not good enough. In conclusion, although there are still a number of optimizations to be addressed, the system presented in this thesis can be considered the first speaker-independent LV-CSR system designed for Romanian.