A Magyar Referencia Beszédadatbázis
és alkalmazása orvosi diktálórendszerek kifejlesztéséhez

VICSI KLÁRA

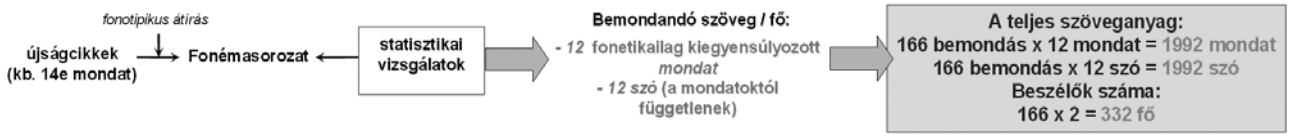
BME Távközlési és Médiainformatikai Tanszék

K+F projektek

Az alábbiakban röviden ismertetésre kerülő projektet az IKTA támogatta (IKTA-00056/2003), a konzorcium tagjai a BME Távközlési és Médiainformatikai Tanszék Beszédakusztikai Kutatólaboratóriuma és az MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoportja voltak. A projekt célkitűzése egy általános magyar nyelvű folyamatos beszédfelismerési technológia kidolgozása volt, valamint egy ahhoz tartozó nyelvi modell elkészítése, amelynek segítségével a rendszer alkalmas orvosi leletek diktálásakor a lelet automatikus lejegyzésére. Az elért eredményeket az alábbi két boxban mutatjuk be.

A Magyar Referencia Beszédadatbázis

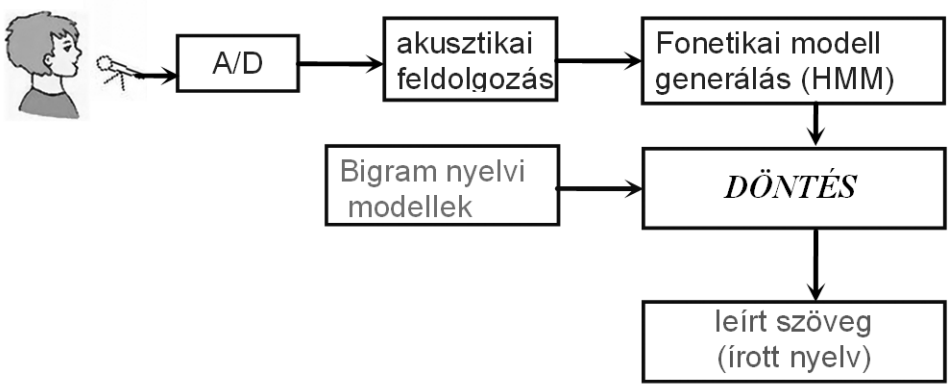
Az adatbázis szöveganyagának megtervezése:



Endoszkópos, pajzsmirigy-scintigráfiás, hasi ultrahang leletek diktálása

Kifejlesztésre került egy Windows XP alatt működő beszédfelismerő fejlesztői környezet, amely alkalmas különböző *középszótáras 1000-10000 szavas* szövegek betanítására és felismerésére.

A felismerő a statisztikai alapon működő HMM akusztikai fonémamodellekkel, valamint a statisztikai alapú bigram nyelvi modellel működik, nemlineáris simítást használva. Az *akusztikai* modelleket az MRBA beszédadatbázissal tanítottuk.



A *nyelvi betanításhoz* a budapesti SOTE II. sz. Belgyógyászati Klinikájától és a Szegedi Orvostudományi Egyetemről gyűjtött korábbi leletanyag korpuszt használtuk. Ezen szövegkorpusz alapján elkészítettük el a *teljes szóalakszótárat*, amely 14331 szót tartalmaz, a kiejtési szótárat és ezek téma szerint osztott kisebb szótárait, valamint a korpusz alapján *morfémaszótárat* is készítettünk, amelynek nagysága 6824 morfémaelem.

A felismerő optimális működését az akusztikai és nyelvi modellek változtatásával állítottuk be. Lényegében a nyelvi modellhez bi-gram modelleket használtunk, de az egyik megoldásban a hagyományos szóalakok az alkotó elemek, a másik megoldásban viszont a morfémák.