Acta Agraria Kaposváriensis (2006) Vol 10 No 1, 113-128
Kaposvári Egyetem, Állattudományi Kar, Kaposvár
University of Kaposvár, Faculty of Animal Science, Kaposvár

Egy pajzsmirigy scintigráfiás leletek diktálására alkalmas rendszer technológiai háttere

Kocsor A., Bánhalmi A., Paczolay D.
MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport, 6720 Szeged, Aradi vértanúk tere 1.

ÖSSZEFOGLALÁS

Az automatikus beszédfelismerési technológiák jelentős fejlődésével számos adminisztrációt megkövetelő szakmában megfogalmazódott az igény az ún. beszédalapú dokumentálásra. Különösen igaz ez az orvosi vizsgálati eredmények rögzítésére, amely folyamat felgyorsítása különösen nagy jelentőséggel bír. Kisebb és speciális nyelvi tulajdonságokkal rendelkező nyelvekre egyelőre nagyon kevés orvosi diktáló szoftver látott ezidáig napvilágot, amely többek között a nyelvi sajátosságokon túl a magas fejlesztési költségeknek tudható be. Szegeden kifejlesztettünk egy magyar nyelv automatikus felismerésére alkalmas magmodult, amelyre különböző speciális diktáló rendszer építhető. A magmodul tartalmazza az un. akusztikai modellt, amely alkalmas a magyar nyelv fonémakészletének felismerésére és reprezentatív módon történő modellezésére. A modell felépítésére két egymástól relevánsan eltérő megközelítést alkalmaztunk. Az egyik a beszédfelismerésben közismert és gyakran alkalmazott Rejtett Markov Modell, a másik pedig a Szegeden kifejlesztett újszerű sztohasztikus szegmentális megközelítés. Mindkét modell felépítéséhez egy nagyméretű, 500 beszélőt tartalmazó beszédkorpuszt használtunk fel, majd tesztadatbázisokon összehasonlítottuk a modulok teljesítményét. A magmodul mellé – a kifejlesztett módszerek alkalmazhatóságát bizonyítandó – kiépítettünk egy Windows-os környezetben használható pajzsmirigy scintigráfiás leletek diktálására alkalmas nyelvi modult, amelyet 9231 írott pajzsmirigy lelet és több mint 2500 szóalak alapján építettünk fel. Ismertetjük a kifejlesztett pajzsmirigydiktáló-rendszer felépítését, az épített nyelvi és akusztikai modellek technikáját, a modellek hatékonyságát jellemző teszteredményeket, továbbá kitérünk a program felhasználási lehetőségeinek és technikájának különböző aspektusaira is.
(Kulcsszavak: folyamatos automatikus beszédfelismerés, diktáló rendszer, HMM, Rejtett Markov Modell, MSD, morfoszintaktikai leíró, nyelvi modell, akusztikai modell, N-gram)

ABSTRACT

Technological background of a speech recognition system for the dictation of thyroid gland medical reports

A. Kocsor, A. Bánhalmi, D. Paczolay
Research Group on Artifical Intelligence of the address: Hungarian Academy of Sciences and University of Szeged, H-6720 Szeged, Aradi vértanúk tere 1.

With the considerable development of speech recognition technologies in several administration-requiring professions the demand for the so called speech-based documentation has grown. This is particularly true in the case of the documentation of medical reports therefore the acceleration of this procedure is of great importance. For smaller languages with special linguistic features few systems for dictating medical reports have been developed so far

which fact can be attributed to linguistic specialties and high development expenses. In Szeged we developed a core module capable of automatic recognition of the Hungarian language on which several domain oriented systems can be built. The core module contains the so called acoustic model, which is suitable for the recognition of the Hungarian phoneme set and representative modeling. For the building of the model we used two significantly different approaches. One is the Hidden Markov Model, well known in speech recognition, the other is the novel stochastic segmental approach developed in Szeged. For the development of both models we used a large speech corpus with 500 speakers, and then the performance of the modules was tested on test databases. To accompany the core module we built a language module (for Windows environment) suitable for the dictation of thyroid gland medical reports in order to justify the applicability of the developed methods. The module was built on 9231 written thyroid medical reports and over 2500 word forms. We present the structure of the developed system for dictating thyroid gland medical reports, the technology of the built language and acoustic models, the test results describing the efficiency of the models, furthermore we mention the different aspects of the application and technology of the software.
(Keywords: continous speech recognizer, ASR, automatic speech recognition, dictation system, HMM, Hidden Markov Model, MSD, Morphosyntactical descriptor, grammar, accoustic model, N-gram)

BEVEZETÉS

Az informatika gyors fejlődése az utóbbi évtizedekben egyre hatékonyabb beszédfelismerő rendszerek létrehozását tette és teszi lehetővé. A cikkünkben betekintést adunk egy magyar nyelvű diktáló rendszer felépítésének elméleti hátterébe és technikai részleteibe. Leírást adunk az egymástól eltérő akusztikai felismerő modellekről és összehasonlítjuk ezeket különböző tesztadatbázisokon. Bemutatjuk a folyamatos beszéd felismeréséhez kötődő problémákat és azokat a megoldási lehetőségeket, amelyek segítségével egy hatékony diktáló rendszer építhető fel. Egy folyamatos beszédet felismerő rendszer alapvető fontosságú része a megfelelő nyelvi modell. Kutatócsoportunk az elmúlt években kutatási célokra kifejlesztett egy általános célú beszédfelismerőt, amelynek most elkészült egy scintigráfiás leletek diktálására alkalmas változata. Az új rendszer segítségével teszteljük le a különböző nyelvi modellek szerepét.

ISMERTEBB DIKTÁLÓ RENDSZEREK

A beszédfelismerő rendszerek nemzetközi piacán mára szinte uralkodóvá vált a Scansoft, amely több felismerő rendszert vásárolt fel és kínál különféle néven. Ilyen termék angol nyelvre a Dragon Naturally Speaking, vagy a IBM ViaVoice (Smith, IBM, 2002). Ezek a rendszerek több éves múltra tekintenek vissza, tapasztalataink szerint igen megbízhatóak a felismerési pontosságot tekintve. A Dragon Naturally Speaking-nek létezik egy speciális, orvosi szakszöveg diktálására alkalmas változata is. A rendszer hét nyelvre vásárolható meg.

A Scansoft piacvezető cég még nem adta ki a magyar változatot, ami azért sem meglepő, mert az agglutináló nyelvekre lényegesen nehezebb felismerőt készíteni. A magyar nyelvben a problémát az okozza, hogy nagyon magas a szóalakok száma a ragozás miatt.

Magyar nyelvre az első „ipari" beszédfelismerőt a Philips készítette el 2004. júliusában, éppen orvosi diktálásra. A Philips közel 10 éve foglalkozik ilyen rendszerekkel, a magyar változat kiadása előtt már 21 nyelvre készített beszédfelismerőt (Nyers, 2004). A szoftvert – igen magas ára miatt – még nem állt módunkban tesztelni, a

leírások alapján hosszabb adaptáció után a szavakra számított felismerési pontosság akár 95%-os eredményt is elérhet.

Kutatócsoportunk közel 10 éve foglalkozik beszédfelismerő rendszerek kifejlesztésével, valamint ehhez kötődő alapkutatásokkal. Mivel kiindulási alapok nem álltak rendelkezésünkre, szükség volt nagy méretű beszédadatbázisok létrehozására, azok felszegmentálására, valamint különböző felismerő és nyelvtani modulok elkészítésére egyrészt a szakirodalom, másrészt a saját kutatásaink alapján.

KÉT ELTÉRŐ MEGKÖZELÍTÉS A BESZÉDFELISMERÉSBEN

Jelenleg a beszédfelismerésnek – mint a Mesterséges Intelligencia kutatás egyik leginkább kutatott ágának – egyik legfontosabb elérendő célja egy általános diktáló rendszer létrehozása, amely a mikrofonba bemondott mondatokat helyesen alakítja át írott szöveggé. Egy ilyen diktáló rendszer létrehozása nagyon sokféle matematikai (*Vapnik*, 1998), informatikai valamint nyelvészeti problémát vet fel. Figyelembe kell venni az adott nyelv fonetikai-akusztikai sokszínűségét, nyelvtani és szórendi sajátosságait. Emellett a másik nagyon fontos problémakör az, hogy a rendszernek fel kell ismernie a bemondott szavakat, illetve azokat a kisebb akusztikai egységeket, amikből a szavak felépülnek.

A közép- és nagyszótáras felismerők mindegyikének gyakorlati megvalósításakor a legkisebb egység, amelyet a rendszernek fel kell ismernie, az a fonéma. A szavak felismerése ezen alkotóelemek felismerésén keresztül valósul meg. A téma kutatása során több különböző gépi tanuló-osztályozó algoritmus (*Duda et al.*, 2001) létrehozására volt szükség a nyelvi alapegységek és az összetettebb struktúrák (szavak, mondatok) felismerésének érdekében (*Huang et al*., 2001). Az ilyen algoritmusoknak két legfontosabb ága a HMM (Rejtett Markov Modell) alapú (*Becchetti and Ricotti*, 2000) illetve a szegmens alapú megközelítés (*Kocsor et al.* 1999).

A két irányvonalban közös, hogy a mikrofonból érkező digitalizált beszédjelből kis időközönként megfelelő méretű mintát veszünk, és minden ilyen kis jeldarabból bizonyos számú jellemzőt vonunk ki, amelyekkel az adott jeldarabot jól jellemezni tudjuk (*1. ábra*). A beszédfelismerésben használt jellemzőkinyerő algoritmusok száma igen nagy, amelyek közül az összetettebbek a hallás és a központi idegrendszer jelfeldolgozásának vizsgálatából származó tudományos eredményeket is figyelembe veszik (*Rabiner and Schafer*, 1978; *Moore*, 1997).

További közös jellemzője a beszédfelismerésben használatos tanuló algoritmusoknak az, hogy a gépi tanulásnak az un. felügyelt tanulás ágába tartoznak. Ez azt jelenti, hogy például az "a" fonéma tanulásakor megfelelő mennyiségű "a" fonéma példánynak kell előre rendelkezésre állnia, maga a tanuló algoritmus ezeket a példákat "tanulja meg". Ezért a tanító adatbázisokat úgy kell kialakítani, hogy a rögzített beszéden kívül tartalmaznia kell az egyes fonémák időbeli helyét a jelen belül. A fonetikai adatbázis ilyen módon történő előállítását szegmentálásnak nevezzük (*2. ábra*). Egy megfelelő minőségű kisszótáras beszédfelismerő szoftver létrehozásához a tapasztalatok szerint több órányi beszédet tartalmazó felszegmentált adatbázisra van szükség.

HMM alapú megközelítés

A HMM (Hidden Markov Model) egy statisztikai alapú tanuló algoritmus, amelyet beszédfelismerésre az 1970-es évek közepétől alkalmaznak, és manapság a módosított változataival együtt a leginkább elterjedt megoldásnak számít (*Becchetti and Ricotti*, 2000). Lényegében arról van szó, hogy minden egyes fonémához létrehozunk egy – általában – három állapotú, balról jobbra elrendezésű rejtett markov modellt (*3. ábra*).

1. ábra

Jellemzőkből álló vektor előállítása a szignálból.

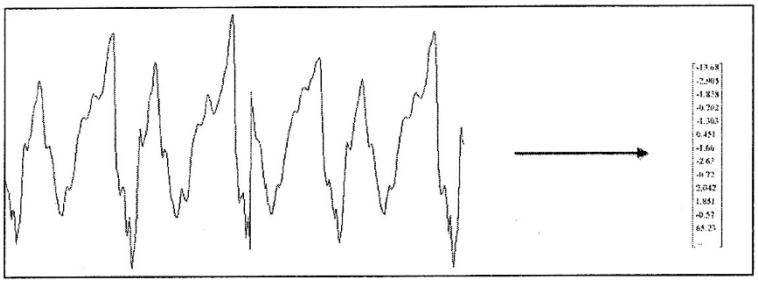
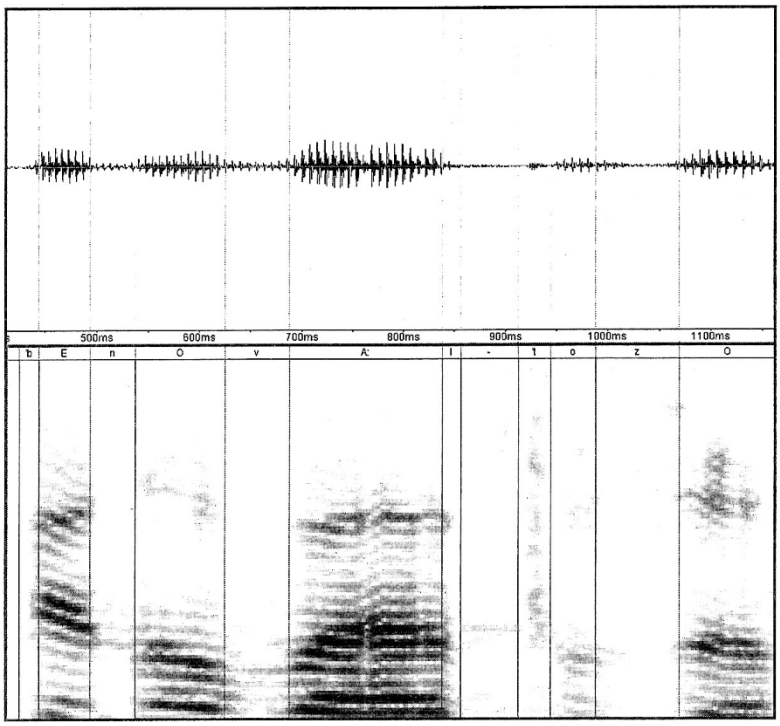


Figure 1: Feature vector extraction from the signal

2. ábra

A szignál fonéma szintű szegmentálása

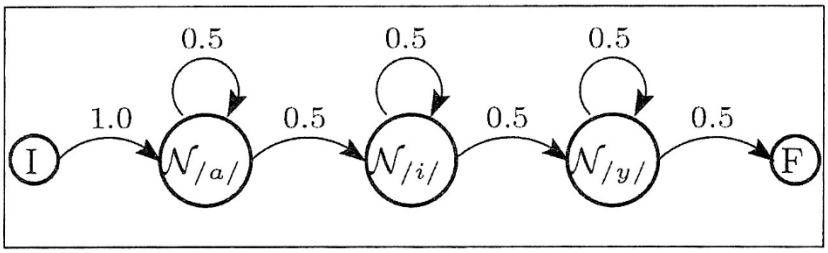


Az ábra felső részén egy szignál darabot látunk, alul a hozzá tartozó spektrumot. A függőleges vonalak a szegmenshatárokat mutatják. *At the top of the picture the signal is shown while at the bottom the corresponding spectrum can be seen. The vertical lines show the borders of the phoneme segments.*

Figure 2: The phoneme level segmentation of the signal

3. ábra

Három állapotú HMM



Három állapotú, balról jobbra topológiájú HMM, fonéma modellezésére. Minden állapothoz hozzá van rendelve egy valószínűségi eloszlás, illetve meg van adva az állapotban maradás és az állapot váltás valószínűsége (lásd nyilak). *Three-state left-to-right HMM for modelling phonemes. All the states have a distribution in the space of the feature vectors and have transition probabilities between states.*

Figure 3: Three-state HMM

Minden egyes modellt betanítunk a megfelelő osztályba tartozó fonéma bemondás példányok segítségével (amiket a szegmentált adatbázisból nyerünk ki). Beszédfelismeréskor a betanított modelleket kiértékeljük a mikrofonból érkező jelből kivont jellemzők alapján. Mindegyik modell egy valószínűségi értéket bocsát ki a jeldarabra, ami azt adja meg, hogy a tesztminta milyen valószínűséggel tartozik a modellnek megfelelő fonémaosztályba. Amennyiben csupán fonémát kívánunk felismerni, akkor azt a fonémát fogadjuk el felismertnek, amelyik modellje a legnagyobb valószínűséget szolgáltatta.

Izolált szavas felismeréskor szükség van egy szótárra, amiben fel vannak sorolva a felismerni kívánt szavak a fonetikus átiratukkal (milyen fonémasorozatból áll össze a szó). Minden egyes szóra fel kell építenünk egy rejtett markov modellt, amit egyszerűen a betanított fonémamodellek összekapcsolásával kapunk (*4. ábra*).

Egy szó bemondásakor a legvalószínűbb modellt választjuk ki.

Folyamatos beszéd felismerése esetében a szavak modelljeit fűzzük össze, így egyre nagyobb és nagyobb struktúrájú HMM-et kapnánk, ami a rendelkezésre álló erőforrásokat gyorsan kimerítené. Ennek kiküszöbölésére elég erős vágásokat alkalmazunk, amiről részletesebben a következő fejezetben írunk.

Szegmens alapú megközelítés

A HMM egyik hiányosságának azt szokás felróni, hogy rosszul modellezi az egyes beszédhangok hosszát. Másik gyenge pontja a fonémák osztályozása, ami más osztályozó algoritmusokkal (ANN, SVM, ...) sikeresebben is megoldható (*Bishop, 1995; Kocsor et al., 2000a; Kocsor et al. 2000b*).

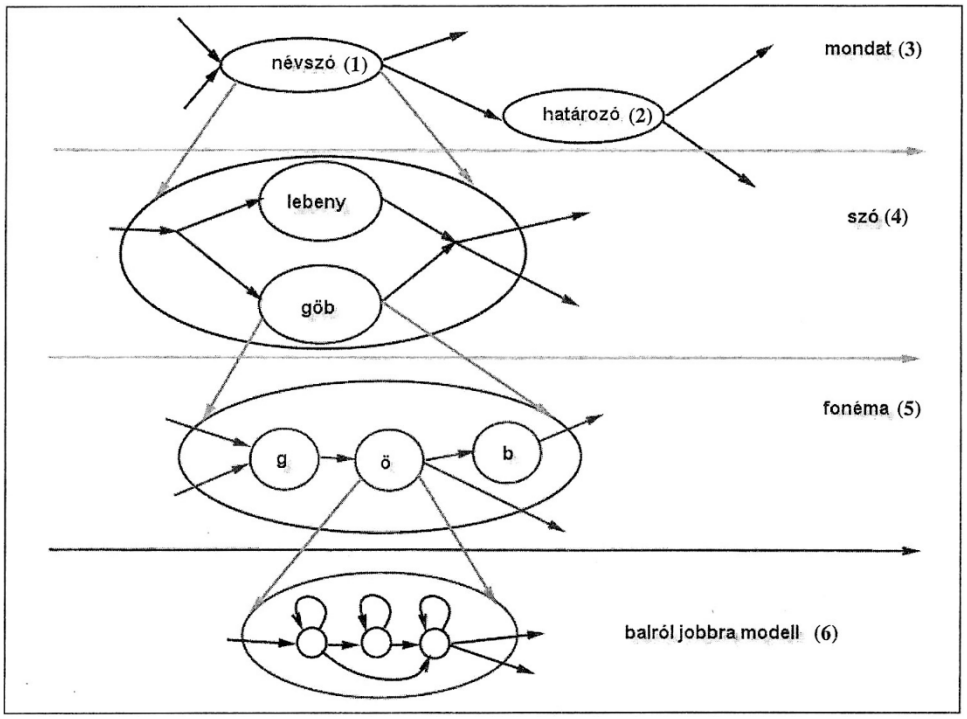
Szegmens alapú megközelítésben a gépi tanuló algoritmus nem egy rögzített kis méretű darabról dönti el adott időközönként, hogy az milyen fonéma része, hanem egy változó időintervallumba eső jelről. Tehát a feladat a legvalószínűbb szegmentálás és fonéma sorozat megkeresése egyidejűleg (Kocsor et al., 1999). Szegmens alapon más jellemzőket is lehet definiálni, például a szegmens hosszát, mely statisztikai úton ügyesebben modellezhető, mint a HMM esetén. A gyakorlatban a lehetséges

szegmenshatárok időbeli pozíciói kvantáltak, elég például milliszekundum pontossággal meghatározni. Ennek az egyik oka az, hogy nem lehetséges a fonémák pontos elkülönítése, mert a folyamatos beszéd során a fonémák folytonosan kapcsolódnak egymáshoz, nincsen közöttük éles határ. A határok bizonytalanságából eredő hibát a sok tanítópélda szerencsére képes ellensúlyozni. A másik oka annak, hogy a határokat csak pár milliszekundumos pontossággal keressük, hogy ez nagymértékben gyorsítja a felismerőt, anélkül, hogy hatékonysága észrevehetően csökkenne. További gyorsítási lehetőség minimális illetve maximális hossz megadása a beszédhangokra (*Tóth et al., 2000*).

További jellemzője a szegmens alapú megközelítésnek, hogy a HMM-nél általánosabb modellnek tekinthető, mert a minimum, maximum és a lépésköz paramétereknek bizonyos beállítása mellett a HMM-nek egy jó közelítését kapjuk meg speciális esetként.

4. ábra

Egy lehetséges HMM modell hierarchia



A hiearchia alsó szintje a fonéma. A HMM fa a fonémától kezdődően épül fel a szavakon át a mondatokig. *The lowest level of the hierarchy is the phonetic level. From phoneme HMMs a tree is built, concatenating the phoneme HMMs to form words and sentences.*

Figure 4: A possible HMM model hierachy

Nominal(1), Adverb(2), Sentence(3), Word(4), Phoneme(5) Left-to-right model(6)

A kétféle irányvonal kísérleti összehasonlítása

A következő fejezetben összehasonlítjuk a HMM és a szegmens alapú akusztikus modelleket. Először ismertetjük az összehasonlításunk módszertani hátterét, majd kiértékeljük a kapott eredményeket.

Az összehasonlítás metodikája, felhasznált beszédadatbázisok

A különböző beszédfelismerő algoritmusok akusztikus modellező képességének összehasonlításakor a következőket szokás elvégezni:

- A fonéma felismerés pontosságának mérése.
- Izolált szó felismerés pontosságának mérése.

Magasabb szintű felismerési hatékonyság mérése (mondatra illetve szövegre) ebben a kontextusban nem szükséges, mivel ez az összehasonlítás inkább az összetett felismerő rendszer, a nyelvtan, a vágás, stb. jóságát méri, és kevésbé a felismerő mag szerepét.

A kiértékelés során fontos, hogy az összehasonlítandó felismerő modulok ugyan-azokon az adatbázisokon legyenek tanítva, illetve ugyanazokon legyenek tesztelve is.

A tanításhoz és teszteléshez a következő adatbázisokat használtuk fel (ezek jellemző statisztikai adatait is feltüntetjük az 1. táblázatban):

1. táblázat

A tanításhoz és a teszteléshez felhasznált adatbázisok

Tanító adatbázis (1)	Fájlok száma (2)	Fonéma szám (3)	Teszt adatbázis (4)	Fájlok száma	Fonéma szám
Szám (5)	2728	20236	Szám	1224	9075
Mtba (6)	3199	154014	Mtba-városok (8)	431	nem szegmentált (9)
Beme-laptop (7)	3315	23762	Beme-laptop	1110	7981
Mtba	3199	154014	Mtba-teszt (10)	687	33159
Timit	3696	142910	Timit	1344	51681

Table 1: Database used for training and testing

Database for training(1), The total number of wave files(2), The total number of phonemes(3) Database for testing(4) Numbers(5), MTBA: Hungarian Telephone Speech Database(6) Our special speech database (7) MTBA cities(8), Not segmented(9), MTBA test(8),

A szám adatbázisok 1 és 1000000 közötti számok bemondásit tartalmazzák. Az Mtba adatbázis különböző telefonon keresztül bemondott szövegek állományait tartalmazza. Ezen belül az Mtba-városok városnevek bemondásaiból áll. A Beme-laptop adatbázis a Beszédmester saját fejlesztésű beszédterápiás szoftverhez készített, férfiak, nők és gyermekek bemondásait tartalmazó adatbázis. A Timit adatbázison kívül az összes adatbázis magyar nyelvű bemondásokat tartalmaz, és teljesen, vagy részben Szegeden készült.

Fonéma felismerés hatékonysága

A fonéma felismerés hatékonyságának összehasonlításkor a felismerő modulokkal szavakat ismertetünk fel, majd összehasonlítjuk a felismert fonémasorozatot a referencia fonémasorozattal. Az általánosan elfogadott összehasonlító eljárás az un. szerkesztési

távolság (Edit Distance). Az Edit Distance (Crochemore and Ryller, 1994) eljárás azt méri, hogy mennyi módosítással lehet megkapni egyik sorozatból a másikat úgy, hogy csak a törés, a beszúrás és az átírás műveleteket használhatjuk. Legyen például a bemondott szó a "scintigram", ekkor a referencia fonéma sorozat (a szabványos SAMPA kódrendszer segítségével megadva): "s 'ts inti+'grOm". A felismerő által jelzett sorozat lehet például "s - 'ts siEm-tiEi+'g O m". Jelen esetben 4 törlés, 1 átírás és 1 beszúrás művelettel kaphatjuk meg az eredeti sorozatból a felismert sorozatot, így a kettő távolsága 6. A teljes tesztadatbázisra vonatkozó hibát két mennyiséggel szokás jellemezni: correction=(N-d-r)/N, illetve accuracy (N-i-d-r)/N, ahol N a referencia sorozatok teljes hossza, i a beszúrás (insertion), d a törlés (deletion), r az átírás (replace) műveletek száma. A 2. táblázatban foglaljuk össze a kétféle irányvonallal kapcsolatos eredményeket:

2. táblázat

A HMM és szegmens alapú fonéma felismerés hatékonysága

Adatbázis (1)	correction	accuracy	insertion	deletion	replace	N
Mtba-teszt (2)	53.67%/55.99%	43.46%/46.93%	3266/3129	3524/3536	12734/11659	34532
Timit	62.38%/63.09%	53.08%/54.20%	4681/4596	4809/4858	14759/14215	51681

Table 2: The efficiency of the segment and the HMM based phoneme recognition

Databases(1), Hungarian Telephone Speech Database Databases used for testing(2)

A táblázatból az olvasható le, hogy a szegmens alapú felismerő valamennyivel jobb felismerési eredményeket ér el fonéma szinten, mint a HMM. A fonémák illetve fonémacsoportok tanulására a szegmens alapú felismerő ANN (Artifical Neural Network, Bishop, 1995) osztályozó algoritmust használt, míg a HMM az alapértelmezett GMM-et (*Becchetti and Ricotti*, 2000).

Izolált szó felismerés hatékonysága

Izolált szavas teszteléskor a felismerőnek nem fonéma sorozatot kell megadnia, hanem el kell döntenie, hogy melyik szó lett bemondva. A lehetséges szavakat egy szótár tárolja. A teszt eredménye a felismerési pontosság, amelyet a 3 táblázatban foglalunk össze.

3. táblázat

A HMM és a szegmens alapú szó szintű felismerés hatékonysága

Adatbázis (1)	HMM	Szegmens alapú felismerő (2)
Szám (3)	99.50%	98.56%
Mtba-városok (4)	93.35%	89.96%
Beme-laptop (5)	98.47%	93.83%

Table 3: The efficiency of the HMM and the segment-based word level recognition

Databases(1), Segment based recognition(2) Numbers(3), MTBA cities(4), Our special speech database(5)

Izolált szó felismerésénél határozottan jobb a HMM alapú felismerő a szegmens alapúnál, csak 1 helyen értek el közel azonos felismerési pontosságot. Ennek az lehet az oka, hogy a szegmens alapú modellek nem mindig találják meg a szegmens határokat, illetve a HMM egy szegmenst több minta alapján értékel ki, így stabilabb működésű. A fonémák illetve fonémacsoportok tanulása az előző bekezdésben alkalmazott tanítással egyezik meg.

A PAJZSMIRIGY LELETEK DIKTÁLÁSÁRA ALKALMAS RENDSZER FELÉPÍTÉSE

A két magmodul felismerési képessége nem tér el lényegesen, de a HMM alapú rendszer folyamatos felismerésre alkalmazható változata könnyebben implementálható. A további vizsgálatainkat ezért csak a HMM alapú akusztikus magmodul használatával végeztük el. Ebben a fejezetben leírjuk a folyamatos felismerés technikai hátterét és a nyelvi modell felépítésének módját.

A nyelvi modell tanításához használt adatbázisok

Az orvosi beszédfelismerő nyelvtani modelljének létrehozásához egy pajzsmirigy szcintigráfiás leletekből álló szövegkorpuszt használtunk. Az írásos vizsgálati anyagokat 1998 és 2004 között rögzítettük. A közel 9000 leletet a különböző formátumokból egy közös szöveges formátumra kellett konvertálnunk, majd több lépéses javítási folyamat következett. A vizsgálatokról készült minden egyes lelet a következő részeket tartalmazza:

a) fejléc
b) klinikai adatok
c) kérdés
d) előző vizsgálata
e) jelen vizsgálata
f) összefoglaló vélemény
g) aláírás

A szövegkorpuszból töröltük a hiányos leleteket az átvizsgálása során. A szöveges adatbázis létrehozásakor az a) és g) részek nem lettek felhasználva. A végleges adatbázis 8546 szövegből áll. A szövegben 2500 szóalak fordul elő (számok és dátumok nélkül). Átlagosan 11 mondatot, és mondatonként 6 szót tartalmaz egy-egy lelet. A mondatonkénti szavak számának eloszlása nem normális eloszlást mutat, ami annak tudható be, hogy a közel 95000 mondat közül mindösszesen 12500 különböző mondatot tartalmazott az adatbázis.

Folyamatos beszédfelismerés gyakorlati megoldása

Egy diktáló rendszertől egyrészt azt várjuk el, hogy a felismert szöveg a bemondottal megegyezzen, másrészt azt, hogy az írott szöveggé alakítást a bemondással szinkronban végezze. Ha a rendszer úgy működne, hogy minden szó után minden szó következhet, és minden lehetséges szósorozatot a felismerés végéig megtartunk, mint hipotézist, akkor a hipotézisek száma rövid idő alatt kezelhetetlen méreteket öltene, a felismerés sebessége pedig messze elmaradna a bemondás ütemétől. Ennek a problémának a gyakorlati megoldásaként az un. Multi Stack Decoding-ot használhatjuk (*Rabiner and Juang,* 1993). Ennek egy lényeges eleme az, hogy minden időpontban valószínűségük alapján rangsoroljuk az aktuális hipotéziseinket (amelyek tulajdonképpen fonémasorozatok). A rangsorolás valamilyen érték, általában a valószínűség alapján történik. Ezután csak azokat a hipotéziseket tartjuk meg, illetve folytatjuk, amelyek a sorban elől állnak. A

legelterjedtebb két ilyen vágási technika az un. N-Best (az első N legjobb megtartása) illetve a Viterbi (a maximális értéktől való eltérés maximálása). Ahhoz, hogy a folyamatos beszédfelismerő megfelelő hatékonysággal működjön, elegendő, de nem túl sok hipotézist kell időről időre megtartani és tovább vinni, el kell kerülni az azonos hipotézisek ismétlődését, és meg kell találni a megfelelő rangsoroló függvényt. A felsorolt feltételeket a nyelvi modell segítségével teljesíthetjük.

Nyelvi modell fontossága

Mint korábban láttuk, alapvetően a felismerők a legvalószínűbb fonémasorozatokat adják vissza. A nyelvtan szerepe e sorozatok szűrése menet közben, aminek célja a hipotézisek számának minél hatékonyabb korlátozása. Erre van lehetőség, hiszen feltételezhető, hogy csak értelmes szavakat kell felismernie a rendszernek. A szűrés azt jelenti, hogy nem következhet minden szó után minden szó, illetve minden fonéma után minden fonéma egy adott időpontban, a nyelvtan leszűkíti a lehetséges következő fonémák és egyben a következő szavak halmazát, és lehetőség szerint megadja azok valószínűségét.

A szó N-gramm

A legegyszerűbb nyelvtan nem más, mint egy szótár, amelyben a lehetséges szavak és azok kiejtései szerepelnek. Az egyszerű szótár alapú rendszerek hátránya, hogy nem tárolnak semmilyen hosszabbtávú információt a felismert előzményről. A szó N-gramm egy olyan módszer, ami ezt a problémát oldja meg részben. Ez egy statisztika, amely megadja, hogy milyen valószínű egy adott szó az N-1 darab előtte álló szó ismeretében. A gyakorlatban csak kis N-re van esély akkora adatbázis összegyűjtésére, hogy jól közelítsük a beszélt nyelvet. Mivel az N-gramm egy véges korpuszból leszámolt statisztika, így az ezen az alapon működő nyelvtan túl szigorú lehet. Ha egy szó-N-es nem szerepelt az adatbázisban, akkor a nyelvtan azt adja, hogy az ilyen bemondás helytelen. Különféle módszerekkel lehet változtatni az N-gramm szigorúságán, például, ha egy kis pozitív e konstanst rendelünk a hiányzó szó-N-esek valószínűségéhez (*Huang et al.,* 2001). Az e érték megválasztása nem triviális, mert, ha túl nagy, akkor rossz sorozatok valószínűbbek lehetnek, mint a helyesek. Egy jobban paraméterezhető rendszert kapunk, ha e értékét nem konstansnak választjuk. Az e értékét más-más szó-N-esekre az ismert szó-gramm valószínűségi értékek segítségével tudjuk megbecsülni. Tapasztalat szerint a kizárólag N-grammokra épülő nyelvtanok kötött szórend esetén jobban működnek, mint szabad szórend esetén. Kötetlen szórendű nyelv használatakor sajnos további probléma az, hogy kis N-re akár olyan köröket is megenged az N-gramm, ami nagyobb N használatakor nem fordulhatna elő. Azonban az N növelésével exponenciálisan kellene növelni a tanító adatbázis méretét, és ezzel párhuzamosan nőne a tárigény is. Sajnos a magyar nyelv szabad szórendűsége miatt mégis modelleznünk kell hosszú távú kapcsolatokat is, ezért más modellekre is szükség van.

MSD kód alapú szabályok

Az MSD kódolás (morfoszintaktikai kódok) minden szóhoz hozzárendel egy vagy több, 8 hosszúságú karaktersorozatot, amelyek a szavak lehetséges mondattani szerepeit írják le. Az MSD kódok alapján létrehozott szabályokat a következőképpen kapjuk meg: a szövegkorpusz mondataiban lévő szavakat lecseréljük azok mondattani szerepét leíró MSD kód első karakterére. A cserével párhuzamosan előállítjuk a szócsoportokat is, amelybe tarozó szavak ugyanabban a nyelvtani szerepben állhatnak a mondatban (pl. jelző, határozó). Mivel egy-egy szónak más-más mondatani szerepe lehet a különböző mondatokban, így a csoportok nem feltétlenül diszjunktak. A folyamat végén minden

mondathoz egy kódsorozatot rendelünk, ezek összessége adja meg az MSD kódos nyelvtani szabályokat (az azonosak összevonása után). A megvalósítás szempontjából két lehetőség nyílik az elkészült nyelvtan felépítésére. Az egyik lehetőség, hogy a csoportokból keletkezett szótárakat egymás után összefűzzük a szabályok alapján, a másik lehetőség, hogy a szótárakat nem helyettesítjük be, csak a kódját tároljuk el. Az utóbbit beágyazott megoldásnak nevezzük.

Az MSD kódos leírásnál a szavak jelentése eltűnik, csak a szavak mondattani szerepe marad meg. Egy szemléletes példa erre, az hogy, ha tanítás során a „piros alma", „sárga banán" is előfordul, akkor a nyelvtan előállítja a „piros banán"-t is. Az ismertetett N-gramm használatával viszont elvárható, hogy az ilyen jellegű problémák csökkennek, mert megmondja, milyen valószínűséggel követik egymást a szavak. A nyelvtan általánosító ereje az e paraméterrel szabályozható.

A hasonulás

Folyamatos beszédben igen gyakori a szavak közötti hasonulás, ami akusztikailag megváltoztathatja a szavak végét vagy elejét. A hasonulás kezelését bonyolítja az is, hogy nem tudjuk, hogy a beszélő tart-e szünetet a szavak között vagy sem, ezért ezeknek a hasonulást leíró szabályoknak, mint alternatíváknak kell megjelenniük a nyelvtanban. A gyakorlati megvalósítás szempontjából az a kevésbé bonyolultabb, ha előre meghatározzuk a hasonulással keletkező fonéma sorozatokat. Ebben az esetben nem alkalmazható a beágyazott nyelvtani megvalósítás. A hasonulást leíró szabályok összetettsége és nagy száma miatt a hipotézisek bővítése közben (azaz nem előre eldöntve, hanem aktuálisan meghatározva) a hasonulás kezelése nem oldható meg megfelelő sebességgel.

Idő és tár elemzések

Korábban láttuk, hogy a felismerés folyamán mindig több hipotézis él. Minél több a hipotézis, annál lassabb a rendszer, illetve szélsőséges esetben a véges hipotézisszám miatt kiszorulnak lehetőségek. A nyelvtan szerepe, hogy minimális számmal növelje a hipotézisek számát. A szótáralapú megoldás kiegészítve szó 2-, vagy 3-grammal, kis memóriaigényű és kellően gyors, de túl sok hipotézist ad, így összességében lassul a felismerés, különösen akkor, ha a szavak száma nagy. Ezért a szótáralapú nyelvtan főként kis méretű problémákon alkalmazható sikeresen. A szabályalapú nyelvtan futtatása lassabb, a felépítése is lassú és nagyobb memóriaigényű, de kevesebb hipotézist generál így közepes méretű problémákra jól alkalmazható.

Amennyiben a szavak száma több ezer, már más megoldást kell keresni, olyat, aminek a memóriaigénye és sebessége is elfogadható, de lehetőleg ne sok hipotézist generáljon. Erről az utolsó fejezetben írunk.

Előrehozott N-gramm kiértékelés

Az N-gramm egyik legnagyobb hibája, hogy akkor kérdezhető le egy szó-N-es valószínűsége, amikor már ismerjük mind az N szót. Ha ezt a valószínűséget előre szeretnénk tudni, akkor már az elején elkülönülnének az azonos fonémasorozatot tartalmazó hipotézisek is, mert más-más valószínűségű szavakhoz tartoznak, és emiatt azonos hipotézisek ismétlődnének, ami persze elkerülendő. Az irodalomban jól ismert ez a probléma (X. Huang et al., 2001); 1-gramm esetén a megoldás az, hogy minden lehetséges fonéma folytatásra olyan valószínűséget ad meg a nyelvtan, amely az ilyen fonémával folytatódó szavak közül a legvalószínűbb szóhoz tartozik. Ezek a valószínűségek előre kiszámolhatók, és alig növelik a memóriaigényt. 2-gramm esetén már egy táblázatot kell kitölteni a különféle előzményekhez tartozó valószínűségek tárolásához. A táblázatok tárolása már nehezen megoldható, mert a szavak számával

arányosan nő a memóriaigény. Általános esetben (N>2) az N-gramm memóriaigénye exponenciálisan nő az előzmény vektor hosszával (N-1). A gyakorlatban a 3-gramm vagy ennél összetettebb nyelvtan esetén már nem lehet előre eltárolni a táblázatot minden egyes fonéma folytatáshoz. Sajnos valós időben sem lehet megkeresni a maximális értékeket a hipotézis generálás közben. Az utolsó fejezetben egy áthidaló megoldást adunk meg erre a problémára.

Teszteredmények a pajzsmirigy adatbázison

Az első tesztben azt vizsgáljuk, hogyan változik a felismerés pontossága MSD kódos nyelvtan használatával. A teszthez véletlenszerűen kiválasztottunk 100 leletet a pajzsmirigy szövegadatbázisból (részletes leírást a *A nyelvi modell tanításához használt adatbázisok* fejezetben találunk). Minden leletet három részletben olvastak be a beszélők. Egy rész több mondatból is állhat, ezeket 5 beszélőtől rögzítettük, minden beszélő 20-20 leletet olvasott be. Az első részbe a „Klinikai adatok", a „Kérdés" és az „Előző vizsgálat" szakasz került. A második részben a „Jelen vizsgálata", a harmadik részben pedig az „Összefoglaló vélemény" bemondások találhatók. A szótárakat és a nyelvtanokat mindhárom részre külön-külön is elkészítettük. A teszteket az „Összefoglaló vélemény" részekhez tartozó szótárral, nyelvtannal, hanganyaggal végeztük el (*4. táblázat*).

4. táblázat

A HMM alapú folyamatos felismerő hatékonysága

Nyelvi modell (1)	correction	accuracy	insertion	deletion	replace
A, szó követ szót (2)	64,18%	52,03%	108	45	273
B, MSD kódos váz (3)	65,20%	63,17%	18	55	254

Table 4: The efficiency of the HMM-based continous speech recognizer on sentences

Language model(1), A: Word by word(2) B: Class N-grams based on MSD codes(3)

A táblázatból látszik, hogy ha kis mértékben is, de az MSD kódos szabályokat tartalmazó nyelvtan jobb eredményt ér el. Részletesebb vizsgálat után arra a megállapításra jutottunk, hogy habár a nyelvtani szabályok segítenek a nem helyes hipotézisek eldobásában, mégis ha rossz szót szúr be a felismerő, mert aktuálisan az tűnik valószínűnek, akkor nincsen esély a mondat helyes befejezésére.

A következő tesztsorozatban azt vizsgáljuk meg, hogy hogyan javít az N-gramm az előző két nyelvtanon. Az e=0 beállítás egy szigorú vágást ad, mert ha egy szónál a nyelvtan „eltéved", akkor a hiba kijavítását maga az N-gramm nehezíti meg. Az $\varepsilon=10^{-5}$ beállítás esetén könnyebben helyreállhat a felismerés szótévesztés után (*5. táblázat*).

A táblázatból kiolvasható, hogy mindkét nyelvtan által elért eredményeken javít a szó-N-gramm. A két nyelvtan eredménye szó-3-gramm és $\varepsilon=10$ beállítás mellet közelíti meg egymást a legjobban.

Az alábbi tesztekben azt vizsgáljuk, hogyan módosul a felismerés pontossága, ha a hasonulási szabályokat is felhasználjuk a nyelvtan készítésekor (*6. táblázat*).

A hasonulási szabályok alkalmazásával az eredmények csak kis mértékben javulnak, mert a hipotézisek számát jelentősen megnövelheti, miközben a vágás mértéke nem változott.

5. táblázat

A HMM alapú folyamatos felismerő hatékonysága mondatokon szó N-gram használatakor

Nyelvi modell (1)	correction	accuracy	insertion	deletion	replace
A + 2 gramm ε=0	79,95%	79,84%	1	151	27
A + 2 gramm $\varepsilon=10^{-5}$	80,51%	75,45%	45	23	150
A + 3 gramm ε=0	80,63%	75,56%	45	27	145
A + 3 gramm $\varepsilon=10^{-5}$	81,64%	76,80%	43	19	144
B + 2 gramm ε=0	80,96%	80,85%	1	63	106
B + 2 gramm $\varepsilon=10^{-5}$	82,54%	82,20%	3	58	97
B + 3 gramm ε=0	86,93%	86,71%	2	30	86
B + 3 gramm $\varepsilon=10^{-5}$	92,11%	92,00%	1	5	65

Table 5: The efficiency of the HMM-based continous speech recognizer on sentences using word N-grams

Language model(1) A and B signs the language model from Table 4.

6. táblázat

A HMM alapú folyamatos felismerő hatékonysága mondatokon szó N-gram és hasonulás használatakor

Nyelvi modell	correction	accuracy	insertion	deletion	replace
A + h	69,03%	65,31%	33	68	207
A + h + 2 gramm ε=0	85,81%	82,32%	31	24	102
A + h + 2 gramm $\varepsilon=10^{-5}$	86,71%	77,81%	79	17	101
A + h + 3 gramm ε=0	87,83%	85,36%	22	24	84
A + h + 3 gramm $\varepsilon=10^{-5}$	88,06%	80,63%	66	17	89
B + h	68,24%	67,00%	11	104	178
B + h + 2 gramm ε=0	91,21%	91,10%	1	58	20
B + h + 2 gramm $\varepsilon=10^{-5}$	93,35%	93,02%	3	41	18
B + h + 3 gramm ε=0	94,14%	94,03%	1	48	4
B + h + 3 gramm $\varepsilon=10^{-5}$	95,15%	95,04%	1	40	3

Table 6: The efficiency of the HMM-based continous speech recognizer on sentences using word N-grams and assimilation rules

Language model (1) A and B denotes the language model from Table 4, here h means that assimilation rules were used.

A *7. táblázat* azt szemlélteti, miként változik az egyes nyelvtanokkal a felismerés pontossága, amikor azok tízszer nagyobb leletadatbázist írnak le. A szóanyag és a nyelvtan a fent ismertetett módon készült, de most 1000 leletet választottunk véletlenszerűen úgy, hogy ezek tartalmazták az előbbi 100 leletet is.

7. táblázat

A HMM alapú felismerő hatékonysága mondatokon különböző nyelvi modellek használata esetén

Nyelvi modell (1)	correction	accuracy	insertion	deletion	replace
C (2)	55,96%	48,54%	66	68	323
C + 3 gramm (3) ε=0	71,05%	65,32%	51	128	129
C + h (4) gramm $\varepsilon=10^{-5}$	72,63%	67,34%	47	119	124
C + h + 3 gramm $\varepsilon=10^{-5}$	72,07%	70,83%	11	134	114
D (5)	57,32%	55,97%	12	112	267
D + 3 gramm ε=0	73,76%	68,47%	47	109	124
D + 3 gramm $\varepsilon=10^{-5}$	74,55%	73,65%	8	82	144
D + h + 3 gramm $\varepsilon=10^{-5}$	74,55%	74,10%	4	107	119

Table 7: The efficiency of the HMM-based continous speech recognizer on sentences using different language models

Language model(1) C: denotes the word by word language model(2) D: denotes the model using class N-grams based on MSD codes(3) h means that assimilation rules were used (4) "3 gramm" means that word 3-gram model was used

A várakozásainknak megfelelően a szavak számának jelentős növelése minden esetben rontott a felismerési pontosságon.

A teszteredményeket úgy foglalhatjuk össze, hogy a szó-N-gramm javít leginkább a rendszeren, míg a hasonulás és az MSD kódos szabályok alkalmazása kevésbé, így ezek még további vizsgálatra szorulnak.

GYAKORLATI ALKALMAZHATÓSÁG

A folyamatos beszédfelismerő demonstrálására egy általános, könnyen használható diktáló rendszert fejlesztettünk ki. Az akusztikai és nyelvi modulok egymástól függetlenül módosíthatók, így könnyen újabb modellekre cserélhetők. A rendszer alapjaiban támogatja a különböző beszédstílusú felhasználókat, hasonló hangú beszélők meghatározott csoportján tanított akusztikai modellekből kiindulva a felhasználók profájlokat hozhatnak létre. Ilyen kiinduló csoportok lehetnek például: felnőtt nő, férfi vagy gyerek. Későbbiekben a felhasználóhoz tartozó akusztikus modell adaptáció segítségével még pontosabbá tehető, személyre szabható. A különféle szakszöveg diktálására alkalmas nyelvtanok külön is telepíthetők a rendszerhez. A profájl a felhasználóhoz rendelt, telepített nyelvi modellek személyre szabását, illetve adaptációját is tartalmazza. A tervezett nyelvi adaptáció az N-gramm súlyok módosítását jelenti, illetve az egyszerűbb nyelvtanok esetén új szót is lehet majd felvenni a szótárba.

A diktáló rendszer egy toolbart jelenít meg a képernyő felső részén. Itt könnyen választhatunk az elmentett profájlok között, illetve kiválaszthatjuk azt, hogy hol jelenjen meg a bediktált szöveg írott változata: a diktáló rendszer tartalmaz egy egyszerű szövegszerkesztőt, de lehetőség van a Microsoft Word szövegszerkesztő programjába is diktálni. Ha úgy kívánjuk, akkor tetszőleges aktív alkalmazásban is képes megjeleníteni a szöveget a diktáló rendszer. A feldolgozás tetszőleges időben felfüggeszthető, illetve lehetőség van korábban rögzített hanganyag szöveggé konvertálására is.

EREDMÉNYEK ÖSSZEFOGLALÁSA, TOVÁBBI KUTATÁSI ÉS FEJLESZTÉSI TERVEK

A demonstrációs célokra kifejlesztett diktáló rendszerünk jelenleg kis, illetve közepes méretű szótárakkal hatékonyan képes valós időben működni. Célunk egy nagyszótáras piacképes rendszer előállítása. A felismerő magmodul hatékonyságának növelése érdekében az adaptáció megvalósítása elkerülhetetlen. Emellett, a felismerési pontosság növelése érdekében, tervezzük a különböző gépi tanuló algoritmusok kombinációjával (Felföldi et al., 2002) történő felismerésnek a megvalósítását, illetve a HMM és szegmens alapú módszerek ötvözését.

Az MSD kódokon alapuló nyelvtanok a jelenlegi formájukban a tanításhoz használt adatbázisok méretének növelésével olyan memóriaigényűvé válnak, hogy egy átlagos teljesítményű PC-n már nem alkalmazhatók. Az sem elég hatékony módszer, amikor a szabályok szerepét 4, illetve 5 szó-N-grammokkal váltanánk fel. Olyan más nagyobb nyelvtani szerkezeteket leíró megoldásokon dolgozunk jelenleg, amelyek kevésbé memóriaigényesek. A szó-N-grammok valószínűségének eltárolásához szükséges memória nagyságrendekkel csökkenthető, ha a szavakat osztályokba soroljuk, és osztály-N-grammot használunk. Ebben az esetben több hipotézist használunk a szükségesnél. Egy lehetséges megoldás az, hogy az osztályokra készítünk egy nagyobb (4-, vagy 5-gramm) és szavakra egy kisebb (2-, vagy 3-gramm) szótár alapú nyelvtant. Így egyben azt is megoldjuk, hogy előrehozottan kiértékelhessük a nyelvtant (lásd 0). Amennyiben garantálható, hogy a csoportok száma kicsi pl. 16, ami MSD kódok első karaktere esetében igaz, akkor lehetőség nyílik minden a nyelvtan által adott – lehetséges fonémához előre megadni a megfelelő valószínűséget. Ezek után a nyelvtannak az osztály-N-gramm része szintaktikai szabályokat, míg a szó-N-gramm része inkább szemantikai szabályokat szolgáltatna. E probléma megoldása több más probléma kivizsgálását is magával hozza, így ez egy későbbi cikkben kerülhet publikálásra.

IRODALOM

Becchetti, C., Ricotti, L.P. (2000). Speech Recognition, John Wiley & Sons LTD, Chichester, England

C.M. Bishop, (1995). Neural Networks for Pattern Recognition, Oxford University Press Duda, R.O., Hart, P.E., Stork, D.G. (2001). Pattern Classification, Wiley

Felföldi, L., Kocsor, A., Tóth, L. (2002). Classifier Combination in Speech Recognition, Conference of PhD students on Computer Sciences, Volume of Extended Abstracts, Szeged, Hungary, 30-31.

Huang, X., Acero, A., Hon, H. (2001). Spoken Language Processing, Prentice Hall, New Jersey

Kocsor, A., Tóth, L., Kuba Jr., A., Kovács, K., Jelasity, M., Gyimóthy, T., Csirik, J. (2000a). A Comparative Study of Several Feature Space Transformation and Learning Methods for Phoneme Classification, International Journal of Speech Technology, 3. 3/4. 263-276

Kocsor, A., Kuba, A., Tóth, L. (2000b). Phoneme Classification Using Kernel Principal Component Analysis, Periodica Polytechnica, 44. 1. 77-90.

Kocsor, A., Kuba, A., Tóth, L., Jelasity, M., Felföldi, L., Gyimóthy, T., Csirik, J. (1999). A Segment-Based Statistical Speech Recognition System for Isolated/Continuous Number Recognition, Proceedings of the FUSST'99, Aug. 19-21, Sagadi, Estonia, 201-211.

Crochemore, M., Ryller, W. (1994). Text Algorithms, Oxford University Press, Oxford
Nyers, Á., (2004). Beszédfelismerés az orvosi dokumentáció korszerűsítésére, IME (Informatika és Menedzsment az Egészségügyben) 3. 5. 39-43.
Moore, B.C.J. (1997). An Introduction to the Psychology of Hearing, Academic Press
Rabiner, L.R., Juang, B.H. (1993). Fundamentals of Speech Recognition, Prentice-Hall, Englewood
Rabiner, L.R., Schafer, R.W. (1978). Digital Processing of Speech Signals, Prentice-Hall, Englewood
Smith, J., IBM, (2002). ViaVoice and Dragon Naturally Speaking XP, ANWALT S.32
Tóth, L., Kocsor, A., Kovács, K. (2000). A Discriminative Segmental Speech Model and its Application to Hungarian Number Recognition, Springer Verlag, TSD'2000, 307-313.
Vapnik, V.N. (1998). Statistical Learning Theory, Wiley

Levelezési cím (Corresponding author):

Kocsor András
MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport
6720 Szeged, Aradi vértanúk tere 1.
Research Group on Artificial Intelligence of the address: Hungarian Academy of Sciences and University of Szeged
H-6720 Szeged, Aradi vértanúk tere 1.
Tel.: 36-62-544-126, Fax: 36-62-425-508
e-mail: kocsor@inf.u-szeged.hu