Szókezdetek automatikus osztályozása spontán beszédben

Bevezetés

A beszédfolyam automatikus, szavaknak vagy néhány szóból álló szócsoportoknak megfelelő szintaktikai egységekre tagolásában bizonyítottan fontos szerepe van a prozódiai jegyeknek, különösen az alapfrekvenciának és az intenzitásnak (Lehiste 1970; Wright–Taylor 1997; Varga 1994; Hunyadi 2002; Varga 1994; Olaszy 2002, 2005; Szaszák 2009). A prozódiai jegyek mellett a magánhangzó minősége is alkalmazható lehet, elsősorban a szótag eleji és a nem szótag eleji szótagok elkülönítésére, másodsorban pedig a szóhatár meghatározására is (Cruttenden 1997; Kohle 1983; Ladefoged–Maddieson 1990; Pennington 1996; Swerts et al. 2007). A szóhatárok automatikus osztályozásában alapegységként a szótagot szokás felhasználni. Egy szó fonológiai jólformáltsága alapvetően attól függ, hogy a szót alkotó szegmentumsorozat kombinációja jól formált-e (Törkenczy 1994). A fonológiai jólformáltságot meghatározó szabályok a hangsorépítő (fonotaktikai) szabályok. A szótag a szegmentumok sajátos szerveződése, amely jellemző az adott nyelvre, és amelyet a szótagok száma is jellemez. A szótagolási szabályok azt határozzák meg, hogy egy szegmentumsorozatot miként kell szótagokra bontani. Mindezek együttesen a szótagszerkezeti szabályok (Siptár–Törkenczy 2000).

A szótag fonetikai meghatározására vonatkozó kísérletek alapvetően két irányzat köré csoportosultak. Az egyik a beszédprodukció felől közelít, és a nyomatékot, a kilégzés erejét tartja a szótagképzés alapjának. A másik a beszédpercepció felől közelít, és a hangzósságot, azaz a hallhatóságot tekinti a szótag alapjának (vö. Laziczius 1963). A nyomatékon alapuló elméleteket a légzési mechanizmus egyre behatóbb ismerete megcáfolni látszik. A hangzósság a beszédhangoknak az a tulajdonsága, amelynek következtében az azonos hangerőn kiejtett beszédhangok különféleképpen hallhatók. Eszerint a szótag különböző hangzósságú elemek együttese. Az elrendeződés pedig lehet olyan, hogy a hangzósság emelkedő, azaz kevésbé hangzós szótagot hangzósabb követ (pl. *tű*), vagy olyan, hogy egy hangzósabb elemet egy kevésbé hangzós követ, ez az eső szótagtípus (pl. *őz*), avagy lehet emelkedő-eső (pl. *drog*, *kert*).

A magyar kötött hangsúlyozású nyelv, a szóhangsúly az első szótagon van (Siptár–Törkenczy 2000). A szavak a beszéd során azonban rendszerint nem elszigetelten jelennek meg, hanem szavak láncolatában. A folyamatos beszéd során megjelenik a kiemelés, és ebben a tekintetben megkülönböztetünk frázishangsúlyt, mondathangsúlyt, szakaszhangsúlyt, tételhangsúlyt stb. (pl. Szende 1976; Gósy 2004). Vannak olyan szavak, amelyek kizárólag más szavakkal együtt fordulhatnak elő a beszédben, így általában nem kapnak önálló hangsúlyt (névelők, névutók, egyes határozószók stb.). Ezek kiszűrésére a beszédtechnológiában rendszerint valamilyen „stopszólistát” szokás készíteni. A beszéd fizikai megvalósulásakor a kiemelés valamilyen akusztikai paraméter mentén jelenik meg, szegmentális vagy szupraszegmentális szinten. Ismeretes, hogy a beszéd észlelése során alapvetően két típusba sorolható akusztikai információt kell feldolgoznunk: a beszédhangok, a hangkapcsolatok és a hangsorok jellemzőire vonatkozó szegmentális információt és az ezekre mintegy ráépülő, beszédhangokon és hangsorokon átívelő szupraszegmentális információt (Gósy 2004). A szupraszegmentális szerkezet a beszédprodukció által létrehozott komplex beszédjelnek az a vetülete, amely az idő, a frekvencia és az intenzitás folyamatváltozásaiként írható le, és amelynek észlelése állandó viszonyításban lehetséges (Markó 2005). A szupraszegmentumok – legáltalánosabban a hanglejtés, a hangsúly, a tempó, a szünet, a ritmus, a hangszínezet és a hangerő – elsősorban tagoló funkciót töltenek be.

A szóhangsúly szerepének meghatározó volta függ a hangsúly által realizált kiemelés mértékétől, valamint attól, hogy redukálódnak-e a hangsúlytalan szótagmagok. A szóhangsúly fonetikailag az alaphangmagasság, az intenzitás és az időtartam kombinációjaként realizálódhat, ez

a módosulás a legtöbb esetben a zöngés részben történik, és nyelvspecifikusan változik (Morton–Jassem 1965).

A hangsúly létrehozását illetően nincsen egységes magyarázat; feltételezik a produkciós erőfeszítést, az erőteljesebb kilégzést, az izomtevékenység fokozását (vö. Malberg 1968; Fox 2000). Fónagy (1967) és később Ladefoged és Maddieson (1996) is azt emeli ki, hogy a hangsúlyos szótag nem feltétlenül jár együtt konstans akusztikai változásokkal, de állandónak ítélték a produkció során jelentkező nagyobb izomaktivitást. A hangsúly észlelése két tényezőn alapszik: az akusztikai paramétereken és a hallgatónak a hangsúly létrehozására vonatkozó saját tudásán. A hangsúly tehát komplex, akusztikai-motoros jellemzőkkel meghatározható jelenség, a hangsúlyélmény pedig a központi idegrendszerben alakul ki (Fox 2000).

Az első szótagi szóhangsúly a magyar nyelvben valamilyen szegmentális és/vagy szuprasegmentális akusztikai jellemző módosulásaként jelenik meg. A beszédfelismerésben már régóta törekszenek arra, hogy a prozódiai információt adaptálják a már működő rendszerekbe, javítva ezzel a felismerés pontosságát (vö. Freij et al. 1990). A prozódiai sajátosságok figyelembevétele több szempontból is segítheti a beszédfelismerést: egyrészt a szóhatárokkal kapcsolatban szolgáltathat információt, másrészt segíthet a szintaktikai szerkezettel kapcsolatos döntésekben (frázishatárok, kiemelés stb.), harmadrészt pedig a lexikai keresési teret szűkítheti azon nyelvek esetében, ahol a hangsúly jelentésmegkülönböztető szerepű. Wang és Seneff (2001) kiemeli, hogy a hangsúlyos szótagok egyfajta „fonetikai megbízhatósági sziget”-et képeznek, mivel e szótagok esetében a beszédhangok fonetikai jellemzői sokkal tisztábban, kiemeltebben vannak jelen.

A szótag lexikális egységként jó definiálható az írott nyelvre. A beszéd akusztikai egységeként azonban az egyes szótagok fonemikus határai függnek a beszélő beszédsebességétől és ritmikai kiejtésétől (Hirata 2004). A szótagtartamokat nehéz magából a beszédjelből kinyerni (vö. Tamburini 2003), még akkor is, ha szótárt hozzunk létre, amely kapcsolódik az automatikus szótag- és szintaktikai elemzőhöz (Sluijter–Heuven 1996). Ez az automata egy önkényes fonetikai átírást kap bemenő jelként, és a legvalószínűbb szótagösszefűzést adja vissza, de a legtöbb esetben különböző eredményekkel jár, amely mutatja a különböző ejtésváltozatokat. A szó szótagjai származhatnak átlagos beszédtempójú beszélőtől, ahol a szótagok arányos időtartamban jelennek meg, de adódhatnak gyors beszélőtől, és ez kifejezetten nehezen szegmentálható szótagokat eredményez, ilyen például a magyarban a „tehát, tát, teát” (Gósy 2009). Ezért a szakirodalom (Tamburini 2003; Teixeira et al. 2001) és saját korábbi tapasztalataink alapján az tűnik célszerűnek, hogy ne a nehezen meghatározható szótagegységeket modellezzük, hanem a szótagmagot, amely a szótag magánhangzóközpontja (és a határai könnyebben meghatározhatók). A magánhangzókon mért prozódiai jegyek erősen korrelálnak a szótag hangsúlyával (Tamburini 2003; Teixeira et al. 2001), érdemes tehát a szótag magját modellezni a szótag hangsúlyosságának eldöntésére (Tepperman–Narayanan 2005).

Az aktuális felhasználás, jelen esetben a hangsúlydetektálás, határozza meg, hogy az egész szótagot vagy csupán a szótagmagot modellezzük. Ez utóbbinak az az oka hogy a magánhangzó időtartamában egyértelműen mérhetők a prozódiai jegyek. A szótagmagok modellezésére közülük az alaphangmagasságot [f0], az energiát és az időtartamot alkalmaztuk.

A mesterséges beszédfelismerésben a szótagot felismerő rendszerek sokszor a már meglévő beszédfelismerők javításának érdekében jönnek létre. Wang és Seneff (2001) például azt vizsgálta, hogy mennyiben javítható a Jupiter rendszerben a beszédfelismerési teljesítmény akkor, ha a modellben figyelembe veszik a hangsúlyos és a hangsúlytalan szótagok közötti különbséget. A Jupiter olyan automata rendszer, amely telefonon keresztül ad időjárási információt úgy, hogy „megérti” a telefonálók kérdéseit. A szerzők a korábbi telefonbeszélgetések adatbázisát felhasználva azt elemezték, hogy mely akusztikai jellemzők azok, amelyek a legjobban alkalmazhatók a hangsúlyos és a hangsúlytalan szótagok elkülönítésére. Az egyedi jellemzőket vizsgálva az amplitúdó mentén volt a legjobb az elkülönítés, a jellemzők kombinációit elemezve pedig az összesített amplitúdó, az

időtartam, a hangmagasság meredeksége és a zöngésség átlagos valószínűsége adta együttesen a legjobb kombinációt. A hangsúlydetekciós modellt a beszédfelismerésre alkalmazva azt találták, hogy az mintegy 5%-kal növeli a felismerés pontosságát. Ez a vizsgálat működő modellen tesztelve támasztotta alá azt, hogy a beszédfeldolgozási modelleknek figyelembe kell venniük a hangsúly szerepét. A folyamatos beszédfeldolgozásban szerepet játszó hangsúlydetekciót vizsgáló más szerzők kevésbé optimisták ennek az alkalmazhatóságával kapcsolatban. Van Kuijk és Boves (1999) a holland Polyphone adatbázist (több ezer beszélőtől telefonon rögzített, spontán alkotott és felolvasott mondatokat tartalmazó adatbázis) felhasználva azt az eredményt kapták, hogy az alkalmazott algoritmus mintegy 70%-os teljesítményt képes elérni a hangsúlyos és a hangsúlytalan magánhangzók osztályozásában. A szerzők a időtartam, az amplitúdó és a spektrális változások különféleképpen normalizált értékeit vizsgálták. Ezek közül a leghatékonyabb mutatónak a teljes energia, azaz a vizsgált magánhangzó teljes időtartamára összesített amplitúdóérték bizonyult.

Xie és munkatársai (2004) egy angol nyelvet tanító számítógépes szoftver megalkotásához vizsgálták a hangsúlyt. Tanulmányukban egy olyan rendszer kifejlesztésének a lehetőségét tanulmányozták, amely a nyelvtanulók által produkált hangsúlymintázatokat elemzi, és adott esetben kijavítja azokat. Egy ilyen szoftver megalkotásának első lépése annak meghatározása, hogy milyen akusztikai jellemzőket kell majd a rendszernek monitoroznia. A szerzők arra az eredményre jutottak, hogy a hangsúly legmegbízhatóbb jelzése az angolban az időtartam és az amplitúdóinformáció kombinációja. Hasonlóan jól jelezte a hangsúlyt a magánhangzó minősége is, ez azonban nem mondható el az alapfrekvenciáról. Bár a vizsgált akusztikai jellemzők alapján a hangsúly detekciója nem volt gyengébb, mint a fentebb bemutatott tanulmányokban, a szerzők szerint a 80–90%-os teljesítmény nem elegendő ahhoz, hogy ezt az algoritmust kereskedelmi forgalomba kerülő rendszerekben alkalmazzák. Felmerül a kérdés, hogy milyen fonetikai paraméterek alapján történhet a szótag automatikus azonosítása, és ez milyen mértékben nyelvspecifikus?

A szókezdetek automatikus detektálásában elsődlegesen a prozódiai szerkezet nyomon követése és az általa hordozott információ kinyerése a cél, azon belül pedig a szó eleji hangsúlyok azonosítása. A jelen vizsgálat célkitűzése az volt, hogy szegmentális és szupraszegmentális jellemzők alapján automatikusan meghatározzuk a szavak kezdőpontját a spontán beszédben úgy, hogy a szó első szótagjának szótagmagját modellezzük. A szó eleji és a nem szó eleji szótagmagokból kinyertük az alaphangmagasságot, az intenzitást és az időtartamot. Hipotézisünk szerint a szó eleji szótagmagok szegmentális és szupraszegmentális jellemzői statisztikai elvű algoritmusokkal osztályozhatók a spontán beszédben. Feltételezzük továbbá, hogy a kombinált paraméterekkel az osztályzás pontossága növelhető.

Anyag, módszer, kísérleti személyek

A vizsgálatban a beszédhangok osztályozásához a BEA adatbázisból (Gósy 2008) 10 fiatal beszélő (5 férfi és 5 nő) spontán narratívját választottuk ki; életkoruk átlaga 25 év volt (a legfiatalabb 22, a legidősebb 30 éves). A hanganyagokat először szakaszszinten jegyeztük le a Praat szoftverrel (Boersma–Weenink 2005). Ezt követően automatikus szegmentálóval (MAUS szoftver: ftp://ftp.bas.uni-muenchen.de/pub/BAS/SOFTW/MAUS) hangszinten annotáltuk. A beszédhangok felismerését statisztikai elven, rejtett Markov-modell alapján végeztük el, amelyre a HTK fejlesztői rendszert (Young 2005) alkalmaztuk. A beszédhangok rejtett Markov-modelljei (HMM) a hangra jellemző vektorok eloszlását adják meg. A rejtett Markov-modell-algoritmus tanításához a beszédadatbázisból származtatott nagy mennyiségű paraméter szükséges. Az akusztikai paraméterek közül a beszédtechnológiában a rövid idejű spektrális burkoló görbe érzeti transzformációján alapuló eljárások váltak be a legjobban (Mihajlik 2010). A Mel-frekvenciás kepsztrális együtthatókra történő átalakításon

alapuló eljárást (MFC) igen széles körben használják (az emberi hallást is modellezi). Léteznek más, például az emberi percepció transzformációján alapuló (PLP) lényegkinyerő algoritmusok is (Mermelstein 1976; Hermansky 1990).

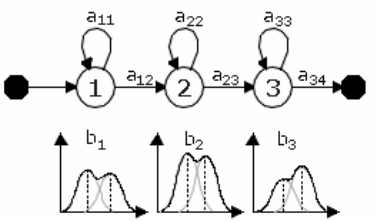
A kutatás során a beszédhangokból kinyertük a spektrális tartalomra utaló Mel-kepsztrális együtthatókat, majd kiszámítottuk ezek első két deriváltját (delták, delta-delták) is. A HMM-ek tanításához a 15 163 beszédhangból a tanításhoz 9985-öt, a teszteléshez 5179 beszédhangot használtuk fel. Az alaphang és az intenzitás vizsgálatához, illetve a szó eleji és a nem szó eleji szótagmagok osztályozásához is a fent leírt korpuszt használtuk (15 163 magánhangzó alaphangmagasságát és intenzitását mértük ki). A glottalizáció miatt azonban 4291 magánhangzót kivettük az adatbázisból, így 10 873 magánhangzón végeztük el az elemzést és az osztályozást. Az alaphangmagasság kinyerésére számtalan módszer létezik. Az általunk használt mód az autokorrelációs függvény maximumának meghatározásán alapuló eljárás (Ainsworth 1976; Iwano 1999). Az eredeti energia- és alapfrekvencia-értékeket 25 ms-os időablakban, 10 ms-os időkeretenként mértük. Az időtartam kinyerése a szótagmagokból viszonylag egyszerű feladat: az automatikus annotálás és a kézi javítás után az annotációból gépi úton kapható meg.

A szó eleji és a nem szótag eleji szótagmagokat szupportvektorgéppel (SVM) és kernelfüggvényként radiális bázisfüggvényt (RBF) alkalmazva osztályoztuk. A szótag eleji és a nem szótag eleji szótagmagok osztályozására úgy alkalmazhatók, hogy a korpusz minden beszédhangjára kinyerjük az akusztikai jellemzőket, majd a tanítóhalmaz értékeivel betanítjuk az osztályozót. Az osztályozó két szabad paraméterét, a C-t és a Gammát háromszoros keresztvalidációval és kimerítő kereséssel optimalizáltuk. A korpuszban a szó eleji szótagmagok száma alacsonyabb volt, ezért a nem szó eleji szótagmagok számát ehhez igazítottuk. Erre azért volt szükség, hogy az algoritmus ne tanuljon rá jobban az egyik csoportra. Az osztályozót különféle akusztikai jellemzővel tanítottuk. Az osztályozásra alkalmazott algoritmusok működésének kiértékelésére és összehasonlítására meghatároztuk az osztályozás pontosságát. A pontosság azt mutatja, hogy az osztályozó algoritmus milyen mértékben azonosítja helyesen a beszédhangokat: $p = tp/(tp + fp)$, ahol a tp (true positive) a helyesen azonosított beszédhangok száma, az fp (false positive) a tévesen osztályozottak száma.

A rejtett Markov-modell. Ez a modell az automatikus beszédfelismerésben széles körben használatos az egyes beszédhangok modellezésére. A beszédhangok modelljein kívül a rejtett Markov-modelles beszédfelismerő fontos tudásbázisa még a szótár – amely megadja, hogy mely szavak milyen beszédhangsorozatból épülnek fel – és az úgynevezett nyelvi modell, amely azt adja meg, hogy adott szókörnyezetet feltételezve mely egyéb szavak előfordulása megengedett, illetve melyik előfordulás mennyire valószínű. A beszédhangok modelljeinek szerepe az akusztikai beszédjel megfeleltetése, leképezése az egyes beszédhangoknak (illetve beszédhangokra). A beszédjelet általában előfeldolgozzuk, például MFC-vel vagy PLP-vel.

A beszédhangok rejtett Markov-modelljei lényegében a hangra jellemző vektorok eloszlását adják meg. Figyelembe véve a jellemző vektorok spektrális származtatását ez tehát frekvenciatartománybeli modellezést jelent. A Markov-modellek leggyakrabban 3 állapotú, balról jobbra felépítésű modellek – utóbbi azt jelenti, hogy a Markov-modell egyes állapotai között átmenet csak balról jobbra lehetséges. Minden állapothoz tartozik egy valószínűségi eloszlást megadó függvény, amelyet statisztikai eszköztárral, leggyakrabban normális eloszlások szuperponálásával becsülünk a modell úgynevezett betanítása során (1. ábra).

A betanítás során ezeknek a függvényeknek a paramétereit becsüljük meg. A beszédfelismerés során a beszédből előállított jellemzővektorokat hasonlítjuk az egyes beszédhangok állapotainak megfelelő eloszlásokhoz. Minél jobban illeszkedik a jellemzővektor egy adott állapot eloszlásához, annál nagyobb súlyt rendel a hozzá kapcsolódó útvonalhoz a dekóder, azaz a tulajdonképpeni beszéd/szöveg átalakító. A Markov-modellek a beszédfelismerőben valójában kettős feladatot látnak el, a jellemzővektorok osztályozása mellett a beszédjelet illesztik is a neki megfelelő beszédhangsorozatra, azaz meghatározzák az egyes beszédhangok kezdő- és végidőpontjait. Úgy is hasz-

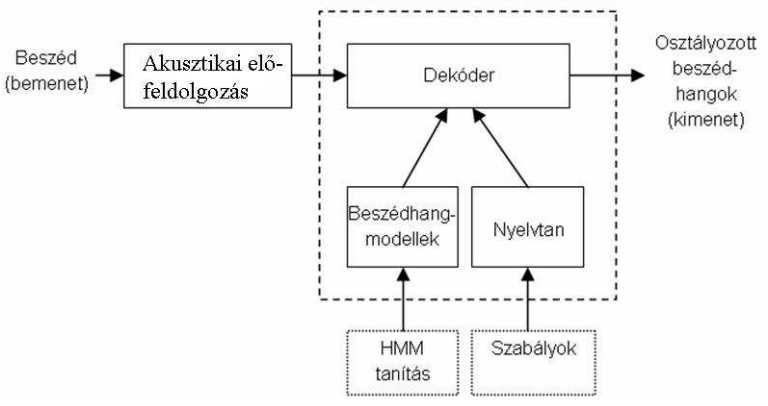


1. ábra

3 állapotú, balról jobbra felépítésű Markov-modell.
A modell tanítás során becsülendő paraméterei az állapotátmeneti valószínűségek (a_{ij}), valamint a normális eloszlások súlya, várható érték- és szórásvektora, amelyek együttesen megadják az eloszlást ($b_j(t)$)

nálhatók, hogy előre elkülönített osztályokra betanított modellek alapján osztályozzanak. Ilyen esetben a beszédfelismerőben használatos szótár szerepét az egyes osztályok listája, a nyelvtan szerepét pedig az osztályozás szabályai veszik át. A beszédhangok osztályozása esetén a lista az osztályozni kívánt beszédhangokból áll, az osztályozás szabályai pedig megadják, hogy milyen beszédhangokat milyen sorrendben lehet illeszteni.

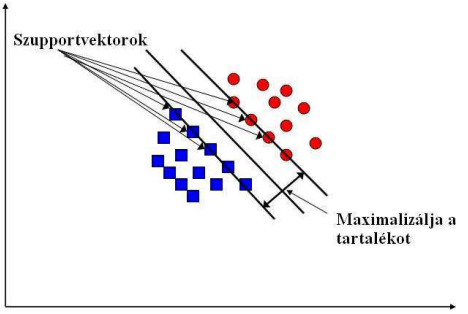
A szótagmagok osztályozása (szó eleji és nem szó eleji szótagmagok) rejtett Markov-modellekkel megvalósítható, a HTK-környezetű osztályozó felépítése a 2. ábrán látható.



2. ábra

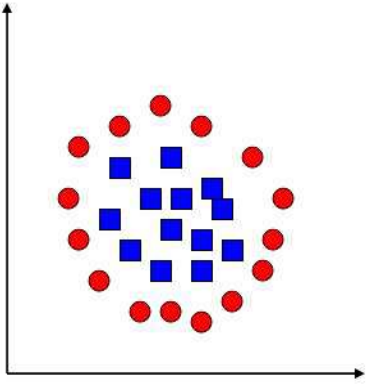
Gépi beszédfelismerő rendszer

Szupportvektorgép (Support Vector Machine). Az SVM olyan matematikai konstrukció, amelyet döntési problémák megoldásához szoktak alkalmazni. Alapverziója a lineáris osztályozók családjába tartozik, de bináris osztályozási problémák megoldására alkalmas. A többi lineáris osztályozóhoz képest az a fő ismérve, hogy nemcsak egyszerűen olyan hipersíkot (más néven vágási síkot) keres, amely elválasztja a pozitív és a negatív tanítómintákat, hanem ezek közül a legjobbat kutatja, vagyis intuitíve azt, amelyik a két osztály mintái között éppen „középen” fekszik (3. ábra).



3. ábra
Két osztály, amely egy hipersíkkal elkülöníthető: lineárisan szeparálható eset

Az SVM tehát olyan döntési hipersíkot határoz meg, amely maximalizálja a tartalékot, azaz a hipersík és a hozzá legközelebbi pozitív és negatív tanítóadatok közti eltérést. Ezeket a tanítóadatokat szupportvektoroknak nevezzük. A hipersík meghatározásában a tanítóadatok közül csak a szupportvektorok játszanak szerepet. Ennek az eljárásnak az előnye egyrészt az, hogy a hipersíkhoz közel álló események osztályba sorolása a legbizonytalanabb; minél kevesebb pont esik erre a területre, annál kevesebb bizonytalan döntést hoz az osztályozó. Másrészt a maximális tartalék által meghatározott szélességű szeparáló sáv elhelyezésére sokkal kevesebb lehetőség van, mint egy tetszőleges szeparáló hipersík esetén. Így kevésbé függ a konkrét adatoktól, ezért az osztályozási modell nagyobb általánosító képességgel rendelkezik. Az SVM-t alapvetően lineárisan szeparálható esetekre találták ki, de előfordulhatnak olyan problémák, amelyek nemlinearitása olyan nagyságrendű, hogy az osztályozó nem lesz hatékony (4. ábra).



4. ábra
Két osztály, amely egy hipersíkkal nem különíthető el: nemlineárisan szeparálható eset

Ennek a problémának a megoldására az adatokat nagyobb dimenziójú térbe transzformáljuk, ahol az adathalmaz már szeparálható. Az erre képes matematikai függvényeket kernel- vagy magfüggvényeknek nevezzük. A magfüggvények segítségével a lineárisan nem szeparálható feladatok szeparálhatóvá tehetők azzal, hogy az adatokat jobban reprezentálható problématérbe transzformáljuk. A gyakorlatban a következő magfüggvényeket szokták alkalmazni: polinominális, radiális bázisfüggvény, kétrétegű perceptron.

A szótag eleji és a nem szótag eleji szótagmagok osztályozására úgy alkalmazhatjuk az SVM-et, hogy a beszédkorpusz minden hangjára kinyerjük az akusztikai jellemzőket (MFCC, PLP, f0 stb.), majd a tanítóhalmaz értékeivel betanítjuk az SVM-osztályozót. A kész osztályozó kiértékeléséhez a teszthalmazt használhatjuk. Vizsgálatunkban az osztályozáshoz az OSU SVM függvénykészletet használtuk (OSU-SVM, MATLAB) az úgynevezett radiális bázis (RBF – Radial Basis Function) kernelfüggvénnyel. Így a szupportvektorgépnek két szabadon állítható paramétere van: C a hibázási paraméter (penalty parameter) és γ az RBF kernelfüggvény (Gauss- függvény) szórásparamétere. Érdemes először egy úgynevezett keresztvalidációs eljárással (cross-validation) és kimerítő kereséssel (grid-search) kizárólag a tanítóhalmazon beállítani az SVM-tanítás említett paramétereit (Hsu et al. 2003).

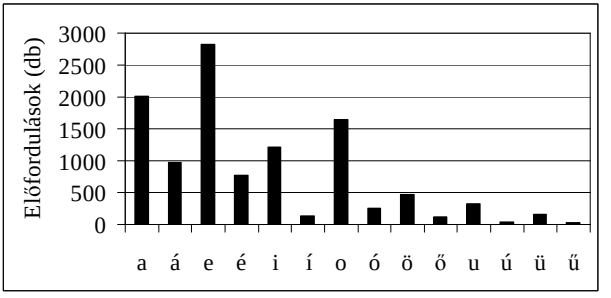
Az SVM RBF kernelfüggvénnyel használva az algoritmusnak tehát két szabadon állítható paramétere van: C és Gamma. Ezek beállítására több módszer is létezik. Az egyik ilyen optimalizációs megoldás az N-szeres keresztvalidáció és kimerítő keresés. Ennek során a tanítóhalmazt véletlenszerűen felbontjuk, jelen esetben három egyenlő részre, majd ezek közül kettőn tanítunk, egyen tesztelünk. A következő lépésben egy másik részen tesztelünk, a többin tanítunk, és ezt annyiszor ismételjük meg, ahány részre a tanítóhalmazt bontották. Az így kapott felismerési arányok átlagát véve becsüljük a felismerési arányokat az SVM aktuális beállításánál. A fentieket elvégezve az SVM számos lehetséges C és γ paraméterpárjára (kimerítő keresés, grid-search) megtalálhatjuk az optimális beállítást, vagyis amikor az SVM a legnagyobb felismerési arányokat éri el. Hsu, Chang és Lin (2003) szerint a C és v értékeket az alábbi tartományokban érdemes keresni:

- C: $\{2^{-5}; 2^{-3}; \ldots; 2^{13}; 2^{15}\}$
- γ: $\{2^{-15}; 2^{-13}; \ldots; 2^{1}; 2^{3}\}$

Eredmények

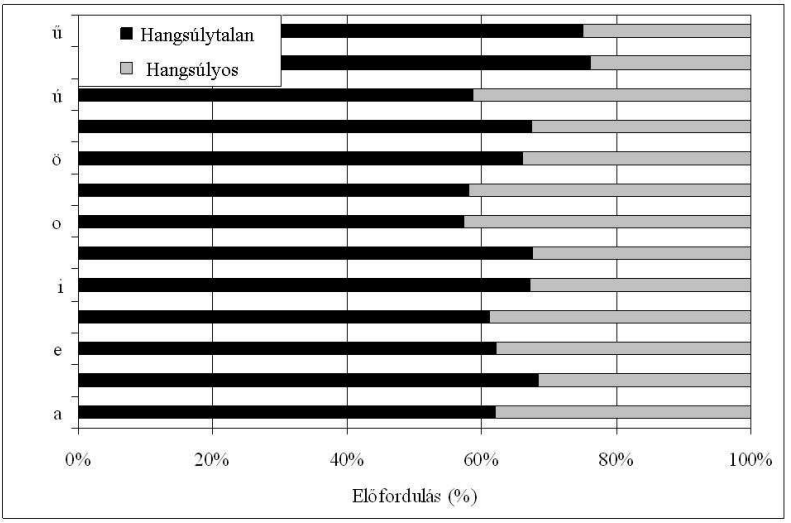
1. A szó eleji és a nem szó eleji szótagmagok jellemzői

A korpuszban a leggyakrabban előforduló magánhangzók az /ɔ/, az /ɛ/ és az /o/ fonémáknak megfelelő beszédhangok voltak. Az /ɛ/ realizációk előfordulása 2823, a legritkább /yː/ realizációké 28 db (5. ábra).



5. ábra
A magánhangzók gyakorisága a korpuszban

Az egyes magánhangzó-minőségeken belül közel azonos arányban fordultak elő a hangsúlyos és a hangsúlytalan magánhangzók. A korpusz magánhangzói (minőségtől függetlenül) átlagosan 34%-a volt található hangsúlyos szótagban (6. ábra).

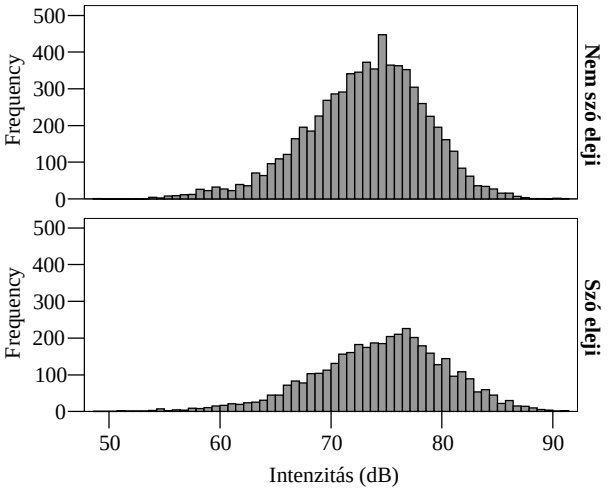


6. ábra
A hangsúlyos és a hangsúlytalan magánhangzók gyakorisága

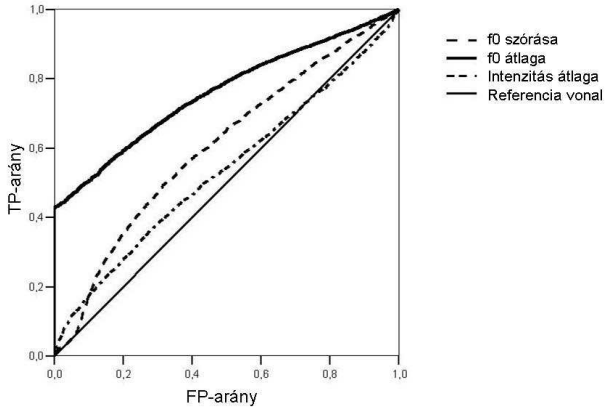
Az átlagos alaphangmagasság értéke a szó eleji szótagmagban 148 Hz volt (szórása 54 Hz), míg a nem szó elejiben 108 Hz (szórása 15 Hz). A különbség szignifikáns (egytényezős ANOVA: $F(1, 8098) = 2004{,}789$; $p < 0{,}001$), ami azt jelenti, hogy a szó elején az alaphangmagasság jellemzően magasabb értéken realizálódik, mint a szó többi részben. A szótagmagban az alaphangmagasság átlagos szórása a szó eleji szótagban nagyobb (7,27 Hz), mint a szó többi szótagjában (5,05 Hz) (egytényezős ANOVA: $F(1, 8098) = 45{,}032$; $p < 0{,}001$). A szó eleji szótagban az alaphang-magasság jellegzetes ingadozást mutat: csúcsra fut, majd a vége felé visszaesik.

Az intenzitás értékében is hasonló tendenciát adatoltunk. A szó eleji szótagmagban az átlagos intenzitás 74 dB volt, a nem szótag eleji szótagmagban átlagosan 73 dB. Jóllehet az átlag nem mutat jelentős különbséget a szótagmag pozíciójától függően, a két csoport között mégis szignifikáns különbség mutatható ki (egytényezős ANOVA: $F(1, 8098) = 22{,}272$; $p < 0{,}001$). Az intenzitás szórása a szótagokban szintén függ a szótag pozíciójától (7. ábra). A szó eleji szótagmagban átlagosan 1,92 dB az intenzitás szórása, míg a nem szó eleji szótagban 2,41 dB (egytényezős ANOVA: $F(1, 8098) = 122{,}89$; $p < 0{,}001$).

Elemeztük, hogy a szó eleji és a nem szó eleji szótagok mely szupraszegmentális jellemző alapján különíthetők el a legjobban. Ennek mérésére ROC (Receiver Operating Characteristic) analízist végeztünk. Ezt a módszert kifejezetten osztályozó rendszerek összehasonlítására szokták alkalmazni, mert megmutatja, hogy egy adott ismertetőjegy milyen átlagos valószínűséggel tud két különböző típusú elemet elkülöníteni a választott küszöbértéktől függetlenül. A ROC-görbének az egyik nagyon fontos mérőszáma a görbe alatti terület; a nagyobb AUC-érték (Area Under Curve) jobb osztályozási értéket jelent (Fawcett 2006). Az egyes jellemzők ROC-analízisből kapott eredményei alapján a legjobb jellemző a szó eleji és a nem szó eleji szótagok elkülönítésében az átlagos alaphangmagasság, amelynek AUC-értéke 0,76. Ezt követi az alaphangmagasság szórásának az AUC-értéke 0,605, majd az intenzitás átlaga (AUC-értéke 0,536), végül az intenzitás szórása, itt az AUC-érték 0,432 (8. ábra).



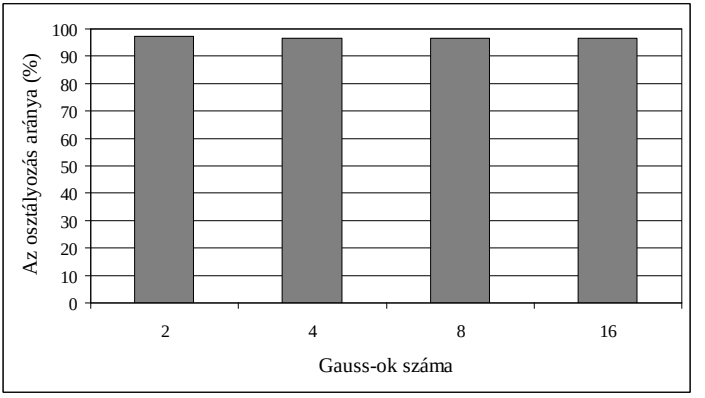
7. ábra
Az átlagos intenzitás a szótag pozíciójától függően



8. ábra
Az egyek akusztikai jellemzők ROC-görbéje (TP-arány = igaz pozitív; FP-arány = hamis pozitív)

A ROC-görbék alapján jól látszik, hogy a legjobb jellemző az átlagos alaphangmagasság és annak szórása, ezt követi az átlagos intenzitás és annak szórása. A szó eleji és a nem szó eleji szótagmagok automatikus osztályozásához elsődlegesen a beszédjelet kell csoportosítani magánhangzókra és mássalhangzókra. Ennek eredményeként azonosíthatók a szótagmagok, és elvégezhető a szótagmagok hangsúlyossági osztályozása. A beszédhangok magánhangzókra és mássalhangzókra történő automatikus osztályozását MFCC-vel előfeldolgozva rejtett Markov-modellel végeztük. Az osztályozni kívánt modellek a következők: magánhangzó, mássalhangzó és szünet. A magánhangzókat, a mássalhangzókat és a szüneteket 3 állapotú HMM-mel modelleztük. A tanítás során „V” szimbólummal jelöltük a magánhangzókat, „C” szimbólummal a mássalhangzókat és „sil”-lel a szüneteket. Mind a három modellt 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítot-

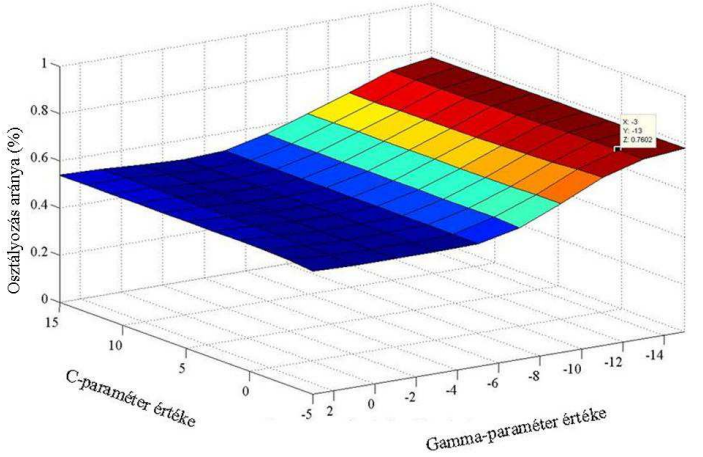
tuk. A nyelvtanban mindhárom hangmodellt („V”, „C”, „sil”) egyenlő súllyal rögzítettük (azaz egyenlő valószínűség mellett). A legjobb felismerési eredményt a 2 Gauss-os modell adta, amely 97%-os volt (9. ábra).



9. ábra
Az osztályozás arányának alakulása a Gaussok számának növelésével

2. A szó eleji és a nem szó eleji szótagmagok osztályozása

A szó eleji és a nem szó eleji magánhangzókat különböző akusztikai paraméterekkel osztályoztuk a spontán beszédben. Először a szupraszegmentális jellemzők alapján végeztük az osztályozást. Közülük az alaphangmagasságot és annak első két deriváltját, valamint az intenzitást és annak első két deriváltját használtuk fel. A tanításhoz az adatok kétharmadát, míg a teszteléshez az egyharmadát használtuk. A 3-szoros keresztvalidáció során akkor volt a legmagasabb a felismerési eredmény, 76%-os, ha a Gamma értéke 2^{-13} és a C értéke 2^{-3} volt (10. ábra).



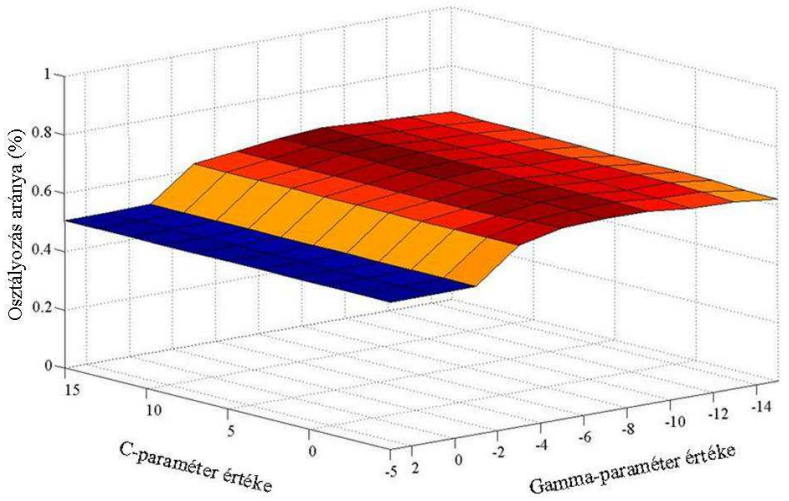
10. ábra
A szótagok osztályozási eredménye a tanításkor 3-szoros keresztvalidáció során

Az ezekkel a paraméterekkel betanított gép a teszthalmazon 78%-os helyes osztályozási eredményt adott. A szó eleji magánhangzókat 70,89%-ban, a nem szó elejieket 80,12%-ban osztályozta helyesen a gép (1. táblázat).

1. táblázat. A szó eleji és a nem szó eleji magánhangzók osztályozásának mátrixa

	Szó eleji	**Nem szó eleji**
Szó eleji	**70,89%**	**29,10%**
Nem szó eleji	**19,87%**	**80,12%**

A második modellben a tanításához és teszteléséhez mfc-ket használtunk és azok első két deriváltját. A tanítás során a legjobb eredmény akkor kaptuk, ha a Gamma 2^{-11}, míg a C-paraméter értéke 2^{15} volt. Ekkor a helyes osztályozás eredménye 66,7% lett (11. ábra).



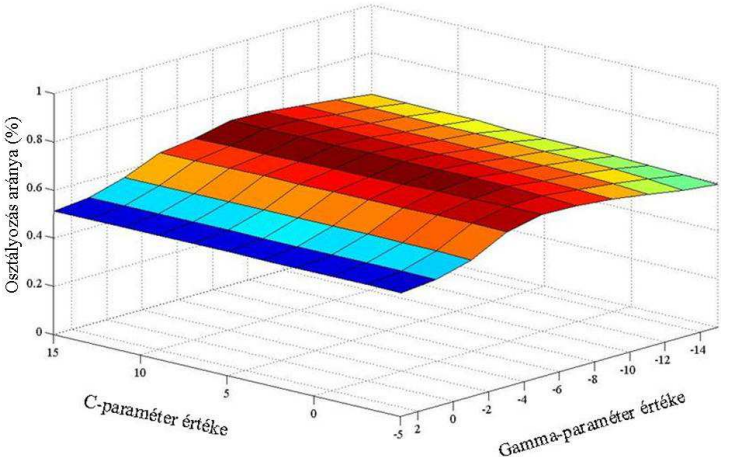
11. ábra
A szótagok osztályozási eredménye a tanításkor 3-szoros keresztvalidáció során

A betanított rendszert a tanításhoz nem használt tesztadatbázison teszteltük, amelynek eredménye 66,03% volt. Az SVM a fenti beállításokkal a szó eleji magánhangzókat 63,3%-ban, míg a nem szó elejieket 68,6%-ban osztályozta helyesen (2. táblázat).

2. táblázat. A szó eleji és a nem szó eleji magánhangzók osztályozásának mátrixa

	Szó eleji	**Nem szó eleji**
Szó eleji	**63,38%**	**36,62%**
Nem szó eleji	**31,32%**	**68,68%**

A harmadik modellben a tanításhoz és a teszteléshez PLP akusztikai jellemzőt és annak első két deriváltját használtuk. A tanítás során 68,5%-os eredményt lehetett elérni. Ekkor a Gamma értéke 2^{-7}, míg a C-paraméter értéke 2^{3} volt (12. ábra).



12. ábra

A szótagok osztályozási eredménye a tanításkor 3-szoros keresztvalidáció során

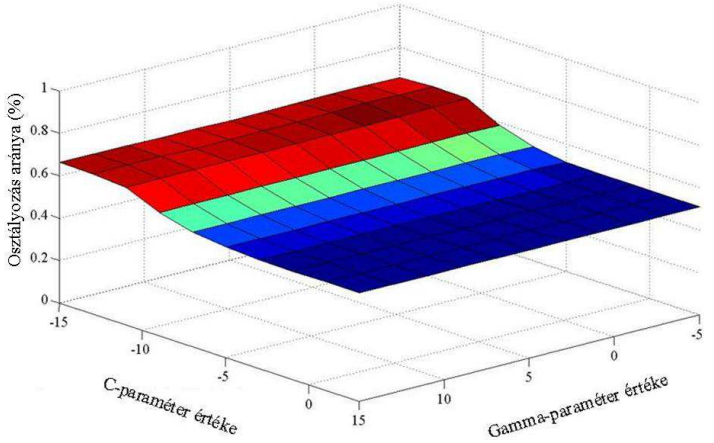
A tanításkor beállított paraméterekkel a teszteléskor 69,29%-os osztályozási arány értünk el. A szó eleji szótagmagokat 62,96%-os, míg a nem szótag elejieket 75,61%-os eredménnyel osztályozta helyesen a gép (3. ábra).

3. táblázat. A szó eleji és a nem szó eleji magánhangzók osztályozásának mátrixa

	Szó eleji	**Nem szó eleji**
Szó eleji	**62,96%**	**37,03%**
Nem szó eleji	**24,37%**	**75,62%**

A negyedik modellben az MFCC- és a PLP-jellemzőket és azok első két deriváltját használtuk a teszteléshez és a tanításhoz. Így egy kombinált jellemzővektort hoztunk létre. A tanítás során a legjobb eredmény 67,14%-os volt, amely gyengébb, mint amit csupán a PLP-jellemzők felhasználásával értünk el (13. ábra).

A teszteléskor a 3-szoros keresztvalidáció során kapott paraméterekkel 66,87%-os helyes osztályozási eredményt értük el. A szó eleji szótagmagokat 57,21%-ban, míg a nem szó eleji szótagmagokat 76,53%-ban osztályozta helyesen az algoritmus (4. táblázat).



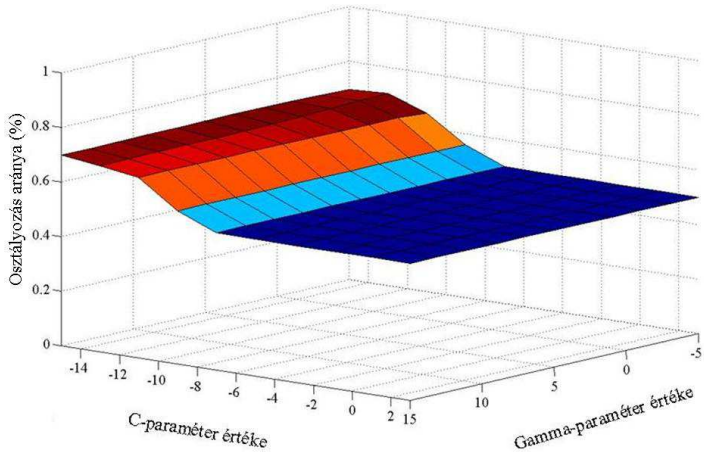
13. ábra

A szótagok osztályozási eredménye a tanításkor 3-szoros keresztvalidáció során

4. táblázat. A szó eleji és nem szó eleji magánhangzók osztályozásának mátrixa

	Szó eleji	**Nem szó eleji**
Szó eleji	**57,21%**	**42,79%**
Nem szó eleji	**23,46%**	**76,53%**

Az ötödik modellben a szegmentális és a szupraszegmentális jellemzőket együttesen alkalmaztuk a tanítás és a tesztelés során. A tanításkor a legjobb eredményt, 70,40%-ot akkor értük el, ha a C-paraméter értéke 2^{-3}, míg a Gamma-paraméter értéke 2^{-13} volt (14. ábra).



14. ábra

A szótagok osztályozási eredménye a tanításkor háromszoros keresztvalidáció során

A kombinált akusztikai jellemzőkkel értük el a legjobb osztályozási eredményt, vagyis 98,9%-ot. A szó eleji szótagmagokat 99,33%-os, míg a nem szótag elejieket 98,48%-os helyes osztályozási eredménnyel ismerte fel a gép (5. táblázat).

5. táblázat. A szó eleji és a nem szó eleji magánhangzók osztályozásának mátrixa

	Szó eleji	**Nem szó eleji**
Szó eleji	**99,33%**	**0,67%**
Nem szó eleji	**1,52%**	**98,48%**

Következtetések

Kutatásunkban arra kerestük a választ, hogy a szó eleji szótagmag és a nem szó eleji szótagmag milyen szupraszegmentális jegyekben tér el egymástól, illetőleg az akusztikai jegyek között milyen hierarchia áll fenn a szeparálóképesség függvényében. Megállapítottuk, hogy mind az alaphangmagasság, mind az intenzitás fontos szerepet tölt be a szó eleji és a nem szó eleji szótagmagok elkülönítésében. A jellemzők közül az átlagos alaphangmagasság mentén válnak el a legjobban a vizsgált szótagok. Az adatelemzésből azonban az is kiderült, hogy sokszor az akusztikai jellemzők csak igen kis mértékben különböznek attól függően, hogy a szótagmag hol fordul elő a szóban.

Az osztályozáshoz első lépésként elkészítettük a spontán beszédben előforduló magánhangzók és a mássalhangzók osztályozására alkalmas statisztikai alapú algoritmust. Ennek eredménye 97%-os volt. A beszédhangokat MFCC-vel előfeldolgozva és rejtett Markov-modellel modellezve jó eredménnyel lehet elkülöníteni a magán- és a mássalhangzóat. Ez az eredmény további felhasználásra ad lehetőséget, amelyet a vizsgálatban a szókezdet automatikus felismerésére alkalmaztuk. A szó eleji és a nem szó eleji magánhangzók automatikus osztályozását szupport vektor géppel és radiális bázis függvénnyel végeztük el. Az SVM-t háromszoros keresztvalidációval és kimerítő kereséssel optimalizáltuk. Az osztályozáshoz különféle akusztikai jellemzőket használtuk. A legjobb eredményt a szegmentális és a szupraszegmentális paramétereket kombináló jellemzővel értük el, ez azt bizonyítja, hogy a szó eleji és a nem szó eleji szótagmagok helyes osztályozásához mind a kétféle tényező jelentősen hozzájárul. A második legjobb eredményt pusztán a szupraszegmentális jellemzőkkel kaptuk. Ez azt valószínűsíti, hogy az osztályozás alapparaméterét ezek a jellemzők adják, és ehhez adódnak hozzá a szegmentális szintű akusztikai paraméterek, amelyek tovább javítják az osztályozás eredményét. A jelen kutatási eredmények hozzájárulhatnak a spontán beszéd automatikus felismerésének megoldásához, mivel ezzel a módszerrel nagy biztonsággal megtalálható a szó kezdőpontja a spontán beszédben.

SZAKIRODALOM

Ainsworth, William Anthony 1976. *Mechanisms of Speech Recognition*. Pergamon Press, Oxford, 1976.

Boersma, Paul – David Weenink 2005. Praat: doing phonetics by computer, http://www.praat.org/ [letöltés: 2005. március 12.].

Cruttenden, Alan 1997. *Intonation*. Cambridge University Press, New York.

Fawcett, Tom 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27: 861–74.

Fox, Anthony 2000. Prosodic features and prosodic structure: the phonology of.suprasegmentals. *Language* 77/3, September 2001.

Freij, G. J. – Fallside, Frank – Hoequist, C. – Nolan, F. 1990. Lexical stress estimation and phonological knowledge. *Computer Speech & Language* 4: 1–15.

Gósy Mária 2004. *Fonetika, a beszéd tudománya*. Osiris Kiadó, Budapest.

Gósy Mária 2008. Magyar spontánbeszéd-adatbázis – BEA. In: Gósy Mária (szerk.): *Beszédkutatás 2008.* MTA Nyelvtudományi Intézet, Budapest, 194–207.

Gósy Mária 2009. Szóejtés és szóészlelés: változatosság és adaptálódás. In: Gósy Mária (szerk.): *Beszédkutatás 2009.* MTA Nyelvtudományi Intézet, Budapest 46–75.

Hermansky, Hynek 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America* 87/4: 1738–52.

Hirata, Yukari 2004. Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics* 32: 565–89.

Hunyadi László: *Hungarian Sentence Prosody and Universal Grammar*. Peter Lang, Berlin, 2002.

Hsu, Chih-Wei – Chang, Chih-Chung – Lin, Chih-Jen 2003. A practical guide to support vector classification. *Technical report, Department of Computer Science*, National Taiwan University. URL http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Iwano, Koji 1999. Prosodic word boundary detection using mora transition modeling of fundamental frequency contours – Speaker Independent Experiments. *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 99).* Budapest, Hungary, vol. 1, 231–4.

Kohle, K. J. 1983. Prosodic boundary signals in. German. *Phonetica* 40: 89–134.

Kuijk, David van − Loe Boves 1999. Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication* 27/2: 95−111.

Ladefoged, Peter – Maddieson 1996. *The sounds of the world's languages*. Blackwell Publishers, Oxford.

Laziczius Gyula 1963. *Fonetika*. Tankönyvkiadó, Budapest.

Lehiste, Ilse 1970. *Suprasegmentals*. Cambridge University Press, Cambridge, Massachusetts.

Malberg, Bertil 1968. *Manual of Phonetics*. North-Holland, Amsterdam.

Markó Alexandra 2005. *A spontán beszéd néhány szupraszegmentális jellegzetessége*. PhD-értekezés. ELTE, Budapest.

Mermelstein, Paul 1976. Distance measures for speech recognition, psychological and instrumental. In: C. H. Chen (ed.): *Pattern Recognition and Artificial Intelligence*. Academic, New York, 374–88.

Morton, John – Jassem, Wictor 1965. Acoustic correlates of stress. *Ls.ng. Speech* 8: 159–181.

OSU SVMs Toolbox for MATLAB (http://www.ece.osu.edu/~maj/osu_svm/).

Olaszy, Gábor (2002): The most important prosodic patterns in Hungarian. *Acta Linguistica Hungarica* 49: 277–306.

Olaszy Gábor 2005. Prozódiai szerkezetek jellemzése a hírfelolvasásban, a mesemondásban, a novella- és a reklámok felolvasásában. In: Gósy Mária (szerk.): *Beszédkutatás 2005*. MTA Nyelvtudományi Intézet, Budapest. 21–51.

Pennington, M. C.: *Phonology in English language teaching: An international approach*. Addison Wesley Longman, New York.

Siptár Péter – Törkenczy Miklós 2000. *The phonology of Hungarian*. Oxford University Press, Oxford.

Sluijter, Agaath M. C. – Heuven, Vincent J. van 1996. Acoustic correlates of linguistic stress and accent in Dutch and American English. *Proc. ICSLP'96*, Philadelphia, 630–3.

Swerts, Marc − Hanne Kloots − Steven Gillis − Georges De Schutter 2007. Vowel reduction in spontaneous spoken Dutch. *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokyo, 31–4.

Szaszák György 2009. *A szupraszegmentális jellemzők szerepe és felhasználása a gépi beszédfelismerésben*. PhD-értekezés. Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Média Informatikai Tanszék.

Szende Tamás 1976. *A beszédfolyamat alaptényezői*. Akadémiai Kiadó, Budapest.

Tamburini, Fabio 2003. Prosodic prominence detection in Speech. *Proc. ISSPA2003*, Paris, 385–8.

Teixeira, Carlos – Franco, Horactio – Shriberg, Elizabeth – Precoda, Kristin – Sonmez, Kemal 2001. Evaluation of speaker's degree of nativeness using text-independent prosodic features. *Proc. Of the Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark.

Tepperman, Joseph – Narayanan, Shrikanth 2005. Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. *Proceedings of ICASSP2005*, 733–6.

Törkenczy Miklós 1999. A szótag. In: Kiefer Ferenc szerk.: *Strukturális magyar nyelvtan 2. Fonológia.* Akadémiai Kiadó, Budapest, 231–4.

Varga László 1994. A hanglejtés. In: Kiefer Ferenc (szerk.): *Strukturális magyar nyelvtan 2. Fonológia*. Akadémiai, Budapest, 468–546.

Wang, Chao – Seneff, Stephanie 2001. Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the jupiter domain. *EUROSPEECH-2001*, 2761–5.

Wright, Helen − Paul A. Taylor 1997. Modelling intonational structure using hidden Markov models. *Paper read at the ESCA workshop on Intonation: Theory Models and Applications*. Athens, Greece.

Xie, Huayang − Peter Andreae − Mengjie Zhang − Paul Warren 2004. Detecting stress in spoken English using decision trees and support vector machines. *ACSW Frontiers*, 145−50.

Young, Steve − Evermann, Gunnar − Gales, Mark − Hain, Thomas − Kershaw, Dan − Moore, Gareth − Odell, Julian − Ollason, Dave − Povey, Dan − Valtchev, Valtcho − Woodland, Phil 2005. *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Departement, Cambridge.

Beke András
ELTE Fonetikai Tanszék
és MTA Nyelvtudományi Intézet

SUMMARY

Beke, András

An automatic classification of word initial sequences in spontaneous speech

Fundamental frequency and intensity have a decisive role in the demarcation of words and syntactic units in continuous speech (Lehiste 1970; Wright & Taylor 1997; Szaszák 2009). We hypothesize that vowel quality may be applicable in the classification of stressed and unstressed syllables of words on the one hand, and in the identification of the onsets of words in spontaneous speech, on the other hand (Cruttenden 1997; Kohle 1983; Ladefoged & Maddieson 1990; Pennington 1996; Swerts et al. 2007).

This study presents a new technique for automatic syllable stress detection. We use various statistical algorithms (SVM: support vector machine; HMM: hidden Markov-model) and various features (f0, intensity, MFCC, PLP) to classify stressed and unstressed syllables.

The results show that stressed and unstressed syllables can be distinguished in 92% of all cases based on the combination of the features analyzed here.

Keywords: spontaneous speech, stressed vs. unstressed syllables, segmental and suprasegmental features, automatic classification