Szótagok automatikus osztályozása spontán beszédben spektrális és prozódiai jellemzők alapján

Beke András[1], Szaszák György[2]

[1] MTA Nyelvtudományi Intézet Fonetikai Osztály
beke.andras@gmail.com
[2] BME Távközlési és Médiainformatikai Tanszék
szaszak@tmit.bme.hu

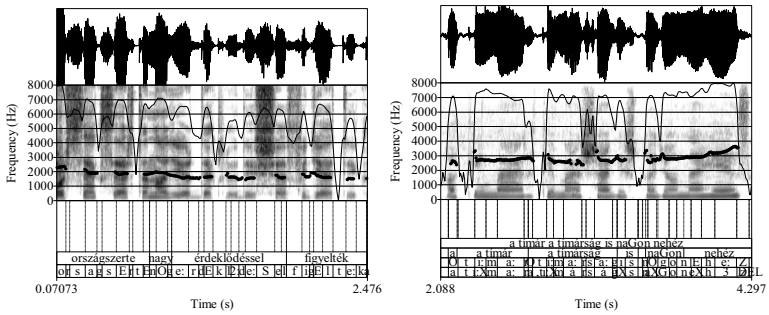
Kivonat: A beszédfolyam automatikus, szavaknak vagy néhány szóból álló szócsoportoknak megfelelő szintaktikai egységekre való tagolásában bizonyítottan fontos szerepe van a prozódiai jegyeknek, az alapfrekvenciának és az intenzitásnak. A prozódiai jegyek mellett a magánhangzó minősége is alkalmazható lehet, elsősorban a szótag eleji–nem szótag eleji szótagok osztályozására, másodsorban pedig a szóhatár meghatározására is. A jelen kutatásban azt vizsgáljuk, lehetséges-e a magánhangzó-minőség alapján a redukálódott magánhangzók automatikus elkülönítése spontán beszédben, illetve magánhangzó-minőség alapján elvégezhető-e a hangsúlyos szótagok automatikus detektálása.

1 Bevezetés

A beszédfolyam automatikus, szavaknak vagy néhány szóból álló szócsoportoknak megfelelő szintaktikai egységekre való tagolásában bizonyítottan fontos szerepe van a prozódiai jegyeknek, az alapfrekvenciának és az intenzitásnak [30]. A prozódiai jegyek mellett a magánhangzó minősége is alkalmazható lehet, elsősorban a szótag eleji–nem szótag eleji szótagok osztályozására, másodsorban pedig a szóhatár meghatározására is [7, 22, 25, 29, 33]. Ha a magánhangzó az eredeti minőségében realizálódik, akkor a hangsúlyos szótag megjelenésének esélye növekszik, míg ha a magánhangzó redukálódott formában realizálódik, akkor a hangsúlyos szótag megjelenésének esélye csökken [17, 33]. A magánhangzó redukciójáról akkor beszélünk, amikor annak képzésekor az artikulációs konfiguráció a centrális irányba tolódik el, megváltoztatva ezzel a magánhangzó minőségét. A jelenség a spontán beszéd esetében fokozottabban nyilvánul meg [1].

Az izolált szavas beszédfelismerésben a szóhatárokat egyértelműen jelzi a szünet jelenléte. A folyamatos felolvasásban a szünet mellett a szupraszegmentális akusztikai jellemzők is hozzájárulnak a szóhatárok pontos gépi meghatározásához. A spontán beszédben azonban a szavak között szinte alig jelennek meg szünetek, a beszéd folyamatos és megakadásokkal tarkított (1. ábra). A korábbi kutatások kimutatták, hogy a humán beszédpercepció szegmentálási eredménye csökken spontán beszédben (felolvasásban 90%-os [2], míg spontán beszédben 70%-os), és az sem egyértelmű, hogy a kísérletben részt vevők milyen akusztikai, szemantikai, pragmatikai jellemzők

alapján jelölték be a megnyilatkozási egységet a szövegben [12]. Az akusztikai kutatások eredményei azt mutatják, hogy a spontán beszédben az artikulációs megvalósítás túlnyomórészt ösztönös, a beszélő nincs feltétlenül tudatában annak, hogy mely szegmentális vagy szupraszegmentális tényezőt alkalmazza tagoló funkcióban, illetőleg meglehetősen nagyok az egyéni eltérések [21].



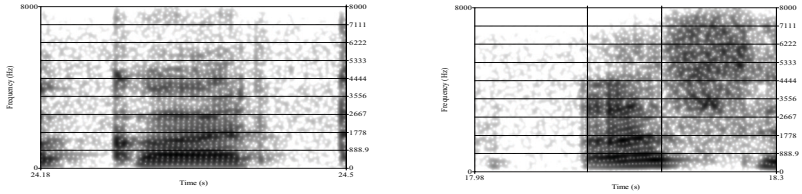
1. ábra. Egy felolvasott mondat részének és egy spontánbeszéd-megnyilatkozás részének akusztikai képe és annotációja.

A gépi felismerés számára a spontán beszédben azonban nem csak a megnyilatkozáshatárok felismerése jelenthet nehézséget, hanem már a szóhatárok bejelölése is. Az egyes idegennyelv-oktatásra irányuló kutatásokban például kimutatták, hogy a szavak szegmentálásának eredménye idegen nyelvben romlik az anyanyelvi szegmentálási eredményekhez képest. A romlás a nem ismert szavak miatt történt [31]. A spontán beszéd gépi felismerésében nagyon nehézkes a szemantikai, pragmatikai jellemzőket beépítése, illetve a szegmentális és szupraszegmentális jellemzők sem egyértelműek, emiatt a spontán beszédben megjelenő szavak határainak meghatározása ezért igen nehéz feladat.

Kutatásunkban arra is keressük a választ, hogy lehetséges-e a szótagok automatikus osztályozása spontán beszédben a magánhangzó minőségét meghatározó spektrális jellemzők alapján. Az osztályozást rejtett Markov-modellel (HMM), valamint szupport vektor gépekkel (SVM) végezzük.

2 Anyag, módszer, kísérleti személyek

A jelen kutatásban a BEA [15] korpuszból 19 magyar beszélő spontán beszédét dolgoztuk fel (8 férfi és 11 nő). Minden hangfájlnak elkészítettük a fonetikus átiratát. Az annotáció során a beszédhangokat kézzel szegmentáltuk a hangszínképük alapján a PRAAT beszédelemző szoftverben. A jelen kutatásban a következő magánhangzókat elemeztük: [ɔ], [aː], [ɛ], [eː], [i], [iː], [o], [oː], [u], [uː]. A beszédhangot akkor jelöltük az annotáció során svá magánhangzónak, (i) ha a beszédhang a centrálishoz közeli formánsstruktúrával rendelkezett (2. ábra), illetve (ii) ha a beszédhang a lehallgatás során svának hatott.



[ɔ] [ə]

2. ábra. Az [ɔ] magánhangzó és a svá [ə] magánhangzó spektrogramja (ugyanazon beszélőtől)

A szegmentálás során az annotációban azt is jelöltük, hogy a magánhangzó hangsúlyos vagy hangsúlytalan szótagon realizálódott. A magyar nyelvben mindig a szó első szótagjára esik a hangsúly [20, 32]. Azt, hogy egy szótag valóban hangsúlyos vagy sem, a szótag F0- és intenzitásértékével ellenőriztük [30].

A svát akkor jelöltük [ə]-val, ha az egységes, osztatlan beszédhangként szerepelt a hangfelismerésben. Nagy karakterrel jelöltük a svá variációkat: [A], [E], [O], ha a svát mint az eredeti magánhangzótól függő realizációt szerepeltettük a modellben. A kutatásban 4000 magánhangzót annotáltunk. A tanításhoz 2500, a teszteléséhez 1500 magánhangzót használtunk. A csoportosításra használt modellek működésének kiértékelésére és összehasonlítására meghatároztuk az osztályozás pontosságát. A pontosság (Acc) azt jellemzi, hogy az osztályozó algoritmus milyen mértékben azonosítja helyesen a beszédhangokat:

Acc=helyesen osztályozott előfordulások / összes előfordulás száma *100%

2.1 A rejtett Markov-modell

A rejtett Markov-modellezéskor az akusztikai előfeldolgozás tekintetében a beszédfelismerésben a beszédhangok akusztikai modellezéséhez használt előfeldolgozási láncot meghagyjuk, és a gyakran alkalmazott MFC (mel-frekvenciás kepsztrális) együtthatókat számítjuk 39 elemű jellemzővektorokat létrehozva. Ezek a jellemzővektorok delta és delta-delta együtthatókat is tartalmaznak. A modellkomplexitást legfeljebb 16 komponenst tartalmazó Gauss keverék (GMM) sűrűségfüggvényig növeljük. Az alkalmazott modellek topológiájuk tekintetében minden esetben 3 állapotú balról-jobbra modellek voltak.

A mintaillesztési megközelítést azonban kissé módosítjuk: felismeréskor nem szósorozatokat illesztünk, hanem beszédhangsorozatokat, az osztályozhatóság szempontjából a vizsgálatunkban érdektelen beszédhangokat azonban töltelékmodellbe vonjuk össze. Ha tehát például egyes magánhangzók és a svá irányába tolódott realizációik elkülönítése (osztályozása) a cél, akkor a magánhangzókra és a svá variánsaikra külön-külön modellek készülnek, minden egyéb beszédhangot töltelékmodellbe vonunk össze, hasonlóan a kulcsszavas beszédfelismeréshez.

A rejtett Markov-modell által időben dinamikus vetemítéssel végzett mintaillesztést kétféle módon valósítjuk meg: az első esetben hagyományos módon a teljes be-

szédmintára illesztünk. Ennek során problémaként jelentkezhet, hogy a töltelékmodellek viszonylagos univerzalitása (általános beszédhangmodell) miatt a mintaillesztés időben pontatlan, különösen hosszú beszédminták esetén. Tapasztalataink szerint ez jelentősen növelheti a törléses-beszúrásos hibák számát. Ha ilyet tapasztaltunk, akkor második mintaillesztési megközelítésként célzottan az osztályozandó magánhangzót tartalmazó részt vágtuk ki közvetlen környezetével együtt, mintegy 100 ms hosszú átmeneti részt meghagyva a magánhangzó előtt és után. Az osztályozás ezután következett, a nyelvi modell azonban kötelező jelleggel 3 modellből álló illesztési szekvenciát engedélyezett csak, amely töltelékmodellel indít és azzal is zárul. A középen elhelyezkedő magánhangzó pedig osztályozandó, az egyes variánsok egyenlő valószínűséggel szerepeltek. Ez a megközelítés kizárja a törléses hibát, és csak helyettesítési hiba fordulhat elő. Ha az osztályozást környezetből kivágottan végezzük, akkor a modellek betanítása is ugyanezen stratégiával, tehát környezetből kivágottan történik. A szerzők a magánhangzók osztályozását hangsúlyos/hangsúlytalan szempontból is vizsgálták ugyanezen megközelítésben. A rejtett Markov-modell alapú osztályozókat HTK környezetben valósítottuk meg [35].

2.2 SVM (Support Vector gépek)

Az SVM (Support Vector Machine) a felügyelt tanulási módszerek családjába tartozik, célja egy olyan szeparáló hipersík keresése, amely jól választja el egymástól a két osztály elemeit (lehet többosztályos is). Az SVM-ek működésének lényege, hogy az eredeti megfogalmazásában még komplex nemlineáris megoldást igénylő feladatot, azaz a feladatból származó mintákat, nemlineáris transzformációk segítségével egy, a bemeneti mintatér dimenziójánál több dimenziós térbe transzformálja, ahol az már lineárisan megoldható. A módszer egyik legnagyobb előnye, hogy egy garantált felső korlátot ad az approximáció általánosítási hibájára. Egy másik fontos jellemzője, hogy a tanulási algoritmus törekszik a modell méretének minimalizálására (ritka modellt alkot), ami a hiba rovására történik, de mértéke egy paraméterrel szabályozható [8, 34]. A hagyományos SVM alkalmazásának legnagyobb akadálya a módszer nagy algoritmikus komplexitása és a nagy memóriaigény, ami tipikusan a nagy adatmennyiség kezelését teszi lehetetlenné. A probléma megoldására számos megoldás született. Ezek az algoritmusok többnyire iteratív megoldások, melyek a nagy optimalizálási feladatot kisebb feladatok sorozatára bontják [3]. A nemlineáris osztályozáshoz a legelterjedtebbet, a radiális bázis (RBF – Radial Basis Function) kernelfüggvényt alkalmaztuk. A hangsúlyos/hangsúlytalan szótagok osztályozását elvégző algoritmus megvalósítása a MATLAB programban történt. Az osztályozáshoz az OSU SVM függvénykészletet használtuk [27]. Az SVM tanításához a magánhangzókból kinyert MFC-jellemzőket használtuk.

3 Eredmények

3.1 A magánhangzó és a redukálódott magánhangzó osztályozása

A semleges magánhangzók akusztikai realizációi jóval változatosabbak, mint a magánhangzókéi [5, 24]. A szegmentális és szupraszegmentális modellekben fontos szerephez juthat a svá automatikus felismerése, hiszen a folyamatos szófelismerésben a magánhangzó nem redukálódott, teljes realizációja jelezheti a szó kezdetét a beszédben [9, 25, 26]. Kopecký [22] a beszédfelismerő rendszerébe beépítette a svá fonémát, amely a rendszer felismerési pontosságának javulását eredményezte.

A magánhangzókat és a svá-realizációkat 3-állapotú HMM-mel modelleztük. A tanítás során "V" szimbólummal jelöltük a magánhangzókat, és "S" szimbólummal a svá-realizációkat. Mind a két modellt rendre 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A nyelvtanban mindkét hangmodellt ("V", "S") egyenlő súllyal rögzítettük (azaz egyenlő valószínűség mellett). A legjobb felismerési eredményt a 4 Gauss-os modell adta (1. táblázat).

1. táblázat: A magánhangzók és a svák felismerési eredményei (4 Gauss).

	Összesen	Acc [%]
"V"	706	79,46
"S"	157	71,97

A semleges magánhangzókra tanított HMM-modell nem veszi figyelembe az eredeti magánhangzót, illetve a szótag hangsúlyosságát. Az eredmények azt mutatják, hogy a spontán beszéden tanított HMM-modellel a svá-realizációk 71,97%-át osztályozta helyesen a rendszer. Az eredményből arra következtethetünk, hogy a semleges magánhangzó rendelkezik egy jól meghatározható spektrális karakterrel, amely megkülönbözteti a többi magánhangzótól. A svá akusztikai realizációi között azonban további kisebb csoportok vannak, amelyek lehetséges svá-alcsoportokra utalnak. Ilyen csoport lehet az, amelyik átfedésben lehet a magánhangzókkal is.

3.2 A magánhangzók és az egységes svá modell

Flemming [10] kimutatta, hogy a svá fonéma realizációinak lehetnek különböző alcsoportjai: közép-centrális svá és kontextusfüggő svá. A svá-realizációk variációinak egy része a kontextus hatására megváltozik, és egy sajátos kontextusfüggő hangminőséget hoz létre, amely a svának egy akusztikai alcsoportja lehet [29]. A nemzetközi és a hazai szakirodalom sem egységes abban, hogy milyen tényleges okai vannak a svá variáltságának, illetve melyek a lehetséges svá-alcsoportok. Flemming szerint nyilvánvaló, hogy a semleges magánhangzónak két típusa létezik, azonban a levonható következtetések nem egyértelműek. A magánhangzó redukciója jelezheti a hangsúlytalan és hangsúlyos szótag közötti szembenállást is, ami az angol nyelvben szabályszerűnek tekinthető. A közép-centrális svá a hangsúlytalan alacsony nyelvállású magánhangzóból jön létre kismértékű redukció során, éppen ezért nem minden magán-

hangzó-minőségből keletkezhet. A svá nem közép-centrális variánsai a magas nyelvállású magánhangzókból jönnek létre a redukció során.

Annak érdekében, hogy a svá-realizációk egységességét megvizsgáljuk, illetve hogy meghatározzuk, melyik magánhangzó minőséghez esik a legközelebb, négy HMM-modellt építettünk. Három modellt készítettünk a három leggyakrabban előforduló magánhangzóra [ɔ, ɛ, o] és egy egységes modellt a semleges magánhangzóra "S". A három magánhangzó-minőséget és a svá-realizációkat 3-állapotú HMM-mel modelleztük. A tanítás során [ɔ], [ɛ], [o] szimbólummal jelöltük a magánhangzókat, és "S" szimbólummal a svá-realizációkat. Mind a négy modellt 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A legjobb felismerési eredményt a 4 Gauss-os modell adta (2.a. táblázat).

2.a. táblázat: Az [ɔ], [ɛ] és [o] magánhangzók, és az egységes svá "S".

	Összesen	Acc [%]
S	140	65
[ɔ]	167	70,65
[ɛ]	225	75,11
[o]	115	73,04

Az eredmények azt mutatják, hogy a svá magánhangzók helyes osztályozásának eredménye 7%-kal romlott ezzel az eljárással. A 2.b. táblázatban a négy modell tévesztési mátrixa mutatja, hogy a svá magánhangzó az [o] modellhez van a legközelebb, mivel az [o] hangok 18%-át téveszti össze a rendszer a svá magánhangzóval.

2.b. táblázat: Az [ɔ], [ɛ] és [o] magánhangzók, és az egységes svá "S" tévesztési mátrixa.

Magánhangzók	[ɔ]	[ə]	[ɛ]	[o]
[ɔ]	118	9	10	16
[ə]	15	91	8	18
[ɛ]	16	13	169	19
[o]	12	7	4	84

A nagyobb tévesztési arány oka az lehet, hogy az [o] vokális artikulációs konfigurációja közel esik a semleges magánhangzóéhoz, illetve az [o] időtartama alacsonyabb, mint az [ɔ] és [ɛ] időtartama [13, 14]. Ennek igazolására kimértük a spontán beszédben előforduló [ɔ], [ɛ], [o] és a redukálódott magánhangzók időtartamát. A svá időtartama szignifikánsan rövidebb, mint a magánhangzóké. A svá magánhangzó időtartama átlagosan 53 ms, míg a magánhangzóké 84 ms (ANOVA: $F(1, 2917) = 252{,}757$;. $p = 0{,}000$**). Ez a tendencia megegyezik a nemzetközi és hazai szakirodalomban leírtakkal [4, 10, 14, 33].

Az adatok szerint a három magánhangzó időtartama közül az [o] magánhangzóé áll a legközelebb a svá időtartamához. Az [o] magánhangzó időtartama (77 ms) szig-

nifikánsan rövidebb, mint az [ɔ] (83 ms) vagy az [ɛ] (90 ms) időtartama (ANOVA: $F(2, 2313) = 19{,}86$ $p = 0{,}000$**; csoportok közötti különbség (post hoc test) $p > 0{,}000$**).

A magánhangzók realizációi átfedésben vannak egymással és a redukálódott magánhangzókkal is az első két formánsértéket tekintve. Bondarko et al. [4] kimutatta, hogy a magánhangzó átmeneti része minden magánhangzó esetében meghatározható mind az olvasott, mind a spontán beszédben, azonban a magánhangzó tisztafázisa a spontán beszédben sokszor eltűnik a magánhangzók redukálódása miatt. A jelen kutatás adatai szerint az [o] magánhangzó időtartama jelentősen rövidebb, mint az [ɔ] és [ɛ] magánhangzóé, ami utal a magánhangzó ejtésekor bekövetkezett célalulmúlásra, ez pedig a magánhangzó tisztafázisának redukciójához vezethet: így az [o] magánhangzó redukálódása olykor erősebb lehet. Eredményeinket alátámasztja Padget [28] kutatása is, amelyben 9 beszélő beszédében a magánhangzók redukálódását vizsgálta. Azt találta, hogy az [ɔ] és az [o] magánhangzót nehezebben lehet elkülöníteni a többi magánhangzótól mind felolvasásban, mind spontán beszédben.

3.3 A magánhangzók és a magánhangzófüggő svá

A helyettesítő funkcióban realizálódott svá akusztikai képe feltételezésünk szerint függ az eredetileg kiejteni kívánt magánhangzó artikulációs konfigurációjától is, amely helyett megjelenik a beszéd során. Ha a svá realizációi függenek a helyettesített magánhangzó minőségétől, akkor a svá-realizációk modellezhetőek a helyettesített magánhangzó minősége mentén. A svá-realizációnak a következő alcsoportjai léteznek: az [ɔ] magánhangzót helyettesítő svá [A], az [ɛ] magánhangzót helyettesítő svá [E], az [o] magánhangzót helyettesítő svá [O]. A tanítás során [ɛ], [o], [ɔ]-val jelöltük az eredeti minőségben realizálódott magánhangzókat, míg [A], [E], [O]-val a helyettesített magánhangzó minőségétől függő svá-realizációkat. Mind a hat modellt 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A nyelvtanban mind a hat hangmodellt egyenlő súllyal szerepeltettük (azaz osztályozáskor egyformán valószínűek voltak). A legjobb felismerési eredményt ismét a 4 Gauss-os modell adta (3. táblázat).

3. táblázat: Az [ɛ], [o], [ɔ] és az [A], [E], [O] osztályozásának eredményei.

	Összesen	Acc [%]
[ɔ]	169	65,08
[A]	47	68,08
[ɛ]	227	69,60
[E]	65	63,07
[o]	116	61,20
[O]	29	62,06

Az eredmények azt mutatják, hogy az osztályozó az [ɔ] magánhangzó helyett realizálódó [A] svát osztályozta a legjobb arányban. Az [o] magánhangzót és az [o] magánhangzó helyett realizálódott svát az algoritmus nem tudta elválasztani olyan pontosan a többi modelltől. Az osztályozás legnagyobb nehézsége ebben az esetben az [o] magánhangzó és a helyette realizálódott svá minőségének variabilitása és az időtartamának csökkenése. A véletlen találgatásnál sokszorosan jobb eredmények mindenesetre alátámasztják, hogy a svá realizációja helyettesítő funkcióban függ a helyettesített magánhangzó eredeti célkonfigurációjától.

3.4 Veláris − palatális magánhangzó és veláris − palatális svá

A magánhangzófüggő svá jobb osztályozhatóságának vizsgálata érdekében megpróbáltunk modelleket összevonni. Korábban megjegyeztük, hogy számos nemzetközi tanulmány foglalkozik a svá-realizációk csoportosítási lehetőségével. A tanulmányok többsége a magánhangzó F2 dimenziójának és időtartamának módosulását tartja a magánhangzó-redukálódás akusztikai paraméterének, ezért a modelleket az F2 dimenzióban vontuk össze. Ha a svá veláris magánhangzó helyett realizálódik, akkor a redukálódott magánhangzóra a veláris magánhangzó-minőség lesz jellemző a svá-realizációkon belül. Ha a svá palatális magánhangzó helyett realizálódik, akkor a redukálódott magánhangzóra a palatális magánhangzó-minőség lesz jellemző a svá-realizációkon belül. Elsősorban a svá-realizációk palatális és veláris alcsoportjainak elkülöníthetőségét teszteltük. Jason [18] a svá lehetséges palatális − veláris alcsoportját HMM modellel tanította és tesztelte fonetikailag variábilis, angol nyelvű, egy beszélőtől származó korpuszon. Eredményei alátámasztották, hogy a svá-realizációknak létezik egy veláris és egy palatális alcsoportja. Azt is kimutatta, hogy a svá magánhangzók kezdeti fázisukban különíthetőek el egymástól, míg a végső fázisukban nem.

A redukálódott magánhangzók realizációiban ugyanúgy létezik veláris − palatális különbség, ahogy a magánhangzók realizációiban. A palatális svá realizációk az F1/F2 térben közelebb vannak a palatális magánhangzókhoz: magasabb F2-értékkel realizálódnak.

A négy magánhangzó-minőséget 3-állapotú HMM-mel modelleztük. A tanítás során VV-vel jelöltük a veláris magánhangzókat, PV-vel a palatális magánhangzókat, VS-sel a veláris svákat és PS-sel a palatális svákat. Mind a négy modellt 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A legjobb felismerési eredményt a 4 Gauss-os modell adta (4.a táblázat).

4.a. táblázat: Az osztályozás eredményei a VV, PV, VS és PS modellekre.

	Összesen [db]	Acc [%]
Veláris mgh.	318	56,91
Palatális mgh.	375	53,86
Veláris svá	76	63,15
Palatális svá	66	40,90

Az eredmények azt mutatják, hogy a svá-realizációk felbonthatóak veláris és palatális svá-realizációkra, amely alátámasztja a formánsok alapján leírt megállapításokat. Az osztályozásban a veláris svá modell adta a legjobb eredményt (63,15%). A palatális svák viszonylag alacsony osztályozási képessége azzal magyarázható, hogy a palatális magánhangzó realizációinak artikulációs tere jóval nagyobb, mint a veláris magánhangzóké, ezért jóval magasabb a realizációk variációinak a száma is (azaz a modell nagyobb szórást enged meg, és emiatt relatíve pontatlanabb modellezést tesz csak lehetővé). Az eredményeinket alátámasztják Bunnel [6] eredményei is: Bunnel megállapította, hogy a palatális svák felismerési eredménye jobb, mint a veláris sváké. Az osztályozási feladatot a veláris és a palatális svá elkülönítésére egyszerűsítve jól látható, hogy a veláris svá felismerése sokkal biztosabb (4.b táblázat).

4.b. táblázat: Az osztályozás eredménye a VS és PS modellekre környezetből kiragadott modellezési technikával.

	Összesen [db]	Acc [%]
Veláris svá	89	79,77
Palatális svá	69	66,66

Ez visszavezethető arra, hogy a veláris svá sokkal egységesebb kategóriát képez, ami megegyezik a nemzetközi szakirodalomban leírtakkal [11]. Harmegnies–Poch-Olivé [16] kimutatták, hogy a redukálódás markánsabban jelenik meg a palatális magánhangzók esetében, mint a veláris magánhangzók esetében.

A nemzetközi eredmények és a jelen kutatás eredményei azt mutatják, hogy a svának helyettesítő funkcióban két alcsoportja különíthető el: a palatális és a veláris svá.

3.5 A modellek kiértékelése

A jelen tanulmányban használt HMM modellek közül az egységes svát és az egységes magánhangzót modellező 3-állapotú HMM-ek pontossága volt a legjobb (78%), 4 Gauss kibocsátási valószínűséget leíró függvénnyel, ami azt jelenti, hogy ezekkel a modelleken osztályozta helyesen a legtöbb hangot az algoritmus (5. táblázat). A legkevesebb helyes találatot az eredeti minőségű magánhangzót és a helyettesített magánhangzó-minőségtől függő svát modellező 3-állapotú HMM-ek adták (69,46%). A veláris és palatális svákat modellező 3-állapotú HMM-ek pontossága 74%.

5. táblázat: A tanított modellek pontossága.

A tanított modellek	Acc [%]
Eredeti magánhangzó-minőség és a helyettesített magánhangzó-minőségtől függő svá	69,46
Veláris palatális magánhangzó és veláris palatális svá (monofon)	70,35
Veláris palatális magánhangzó és veláris palatális svá (környezetből kiragadva)	74,05

Egységes svá és magánhangzók [ɛ],[o], [ɔ]	75,86
Egységes magánhangzó és egységes svá	78,09

3.6 Hangsúlyos − hangsúlytalan szótagok osztályozása a magánhangzó minőségének segítségével

A hangsúlyos és hangsúlytalan szótagok osztályozásához gondosan felszegmentált anyagon, 3-állapotú HMM-eket tanítottunk. A hangsúlyos szótagokat „XA”-val, a hangsúlytalan szótagokat „XT”-vel jelöltük. Mind a két modellt 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk. A nyelvtanban mind a két hangmodellt egyenlő súllyal szerepeltettük (azaz egyenlő valószínűség mellett). A legjobb felismerési eredményt a 8 Gauss-os modell adta (6.a. táblázat).

6.a. táblázat: A XA és a XT osztályozásának eredményei.

Szótagok	Összesen [db]	Acc [%]
XA	309	82,80
XT	855	70,72

Az eredmények szerint a hangsúlytalan szótagok osztályozása kevésbé pontos, ami arra utal, hogy ez az osztály nem egységes a modellezett szótagok hangminősége szempontjából.

Az SVM-mel tanított és tesztelt magánhangzó-minőségen alapuló osztályozó pontossága 54%. A szókezdő pozícióban lévő magánhangzókat mindössze 56%-ban, míg a nem szókezdő pozícióban lévő magánhangzókat 58%-ban osztályozta helyesen az algoritmus (6.b. táblázat).

6.b. táblázat: A szókezdő és nem szókezdő pozícióban lévő magánhangzók osztályozási eredménye (Acc) SVM-mel.

	Szókezdő	Nem szókezdő
Szókezdő	61%	39%
Nem szókezdő	42%	58%

Az SVM-mel végzett osztályozásban is a szókezdő pozícióban lévő magánhangzók eredménye jobb.

A hangsúlytalan szótag modellezésében ronthatja az osztályozási eredményeket, hogy a hangsúlytalan szótagban a magánhangzó minősége nem egységes. A hangsúlytalan szótagban a magánhangzó megjelenhet redukálódott magánhangzóként, illetve az eredeti magánhangzó artikulációs konfigurációnak megfelelő minőségben is. Ennek igazolására három HMM-et építettünk. A kísérlet során modelleztük a hangsúlyos szótagban realizálódott magánhangzót (XA), a hangsúlytalan szótagban realizálódott eredeti magánhangzó-minőséghez közeli magánhangzót (XT) és a hangsúlytalan szótagon megjelenő redukálódott magánhangzót (XS). Ezeket a magánhangzó-minőségeket 3-állapotú HMM-ekkel modelleztük. Mind a három modellt ismét rendre 2, 4, 8, 16 Gauss kibocsátási valószínűséget leíró függvénnyel tanítottuk.

Az osztályozás nyelvtanában mind a három hangmodellt egyenlő súllyal rögzítettük (azaz egyenlő valószínűség mellett). A legjobb felismerési eredményt a 8 Gauss-os modell adta (6.c. táblázat).

6.c. táblázat. A három magánhangzó-minőség (XA,XT,XS) osztályozási eredménye.

Magánhangzó-minőségek	Összesen [db]	Acc [%]
XA	299	80,70
XT	646	68,80
XS	218	73,20

A modellek átlagos osztályozási pontossága lényegesen nem változott az előbbi esethez képest, a hangsúlytalan szótagok helyes osztályozása (közösen az XT- és az XS-modell) ismét közel 71%-os. A hangsúlytalan modell kettéválasztása az eredeti hangminőséghez közeli magánhangzóra és redukálódott magánhangzóra viszont azt mutatja, hogy ha a magánhangzó redukálódik is, akkor valamelyest kisebb eséllyel osztályozza az osztályozó hangsúlyosnak. Meg kell jegyeznünk, hogy a hibák nagyobb része törlésből és nem tévesztésből származott, ilyenkor az osztályozó egyszerűen átugrik egy szótagot (vagy ha úgy tetszik, összevonja azt az előtte-mögötte állóval).

4 Következtetések

A jelen tanulmány célja az volt, hogy a helyettesítő funkcióban lévő svá-realizációkat spektrális jellemzőik alapján modellezze HMM-ekkel magyar nyelvű spontán beszédben.

Az elemzések során bemutattuk, hogy (i) a [ə] és a svá-realizációk MFC-együtthatók alapján előfeldolgozva HMM-ekkel modellezhetőek a magyar nyelvű spontán beszédben; (ii) a svá-variációk realizációi függnek az általuk helyettesített magánhangzó artikulációs konfigurációjától. Ezen megállapítások igazolására hat különböző modellt építettünk, amelyek reprezentálták a „svá" és a lehetséges svá-alcsoportok realizációit. Jóllehet az osztályozás során globális pontosság szempontjából a legjobb eredményt az osztatlan svá, az eredeti minőségben realizálódott magánhangzó modellhalmaz adta, a svá-realizációk közötti leghatékonyabb osztályozást a veláris és palatális svá-alcsoportra épített HMM-ek adták (79%).

A vizsgálat során összehasonlítottuk az eredeti minőségben realizálódott magánhangzók és a redukálódott magánhangzók akusztikai szerkezetét (időtartam és formánsszerkezet). Az eredmények azt mutatták, hogy az [o] magánhangzó artikulációs konfigurációjában közelebb áll a svá artikulációs konfigurációjához a spontán beszédben, mint a vizsgálatban szereplő többi magánhangzó. Ennek oka az, hogy az [o] artikulációs konfigurációja és időtartama jóval nagyobb variációt mutat, mint a vizsgálatban szereplő többi magánhangzóé.

A svá realizációk lehetséges alcsoportjait HMM-ekkel modelleztük. A hipotézisünk és a nemzetközi szakirodalom szerint a svá-realizációk alapvetően két csoportra bonthatók, méghozzá palatális és veláris svá-variációkra. Az eredmények azt mutatták, hogy a svá-realizációknak ez a két alcsoportja létezik: veláris és palatális svá. Ennek oka az, hogy a helyettesítő funkcióban lévő svá függ az általa helyettesített magánhangzó minőségétől: a veláris magánhangzó redukálódása közben megőrzi az alapvető veláris spektrális jegyeket, ahogy a palatális svá is megőrzi a palatális magánhangzó spektrális jellemzőit. Ez természetesen nem zárja ki azt, hogy a svá-realizációk esetleg más dimenziókban is elválaszthatók egymástól.

A hangsúlyos és hangsúlytalan szótagokban realizálódott különböző magánhangzó-minőségeket HMM-ekkel modelleztük MFC-előfeldolgozás alapján. A hangsúlyos és a hangsúlytalan szótagok osztályozásának eredménye 73,76% volt. A hangsúlyos szótagok felismerése ezen belül 82,8%-kal a leghatékonyabb volt. A svá-modell beépítése összességében javított a hangsúlyos és hangsúlytalan szótagok csoportosításában. A magánhangzó-minőséggel modellezett hangsúlyos és hangsúlytalan szótagok felismerésének eredménye jobbnak bizonyult a hasonló nemzetközi kutatások eredményeihez képest (vö.[19, 23]).

Bibliográfia

1. Beke A.: A veláris magánhangzók stabilitása a spontán beszédben. In: Gecső Tamás − Sárdi Csilla (szerk.) A kommunikáció nyelvészeti aspektusai. Kodolányi János Főiskola−Tinta Kiadó, Székesfehérvár−Budapest (2009) 27–31
2. Batliner, A, Kompe, R., Kießling, A., Mast, M., Niemann, H., Nöth, E., Oth, E.N.: M=Syntax+Prosody: a syntactic-prosodic labelling scheme for large spontaneous speech databases. Speech Communication Vol. 25. (1998) 193–222
3. Bennett, K. P., Campbell, C.: Support Vector Machines: Hype or Hallelujah?. SIGKDD Explorations Vol. 2 No. 2 (2000) 1–13
4. Bondarko, Liya V., Volskaya, Nina B., Tananaiko, Svetlana O., Vasilieva, Ludmila A.: Phonetic properties of Russian spontaneous speech. In: Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona, 3-9 August (2003) 2973−6
5. Browman, C. P., Goldstein L.: Articulatory phonology: An overview. In: Phonetica Vol. 49 (1992) 155−80
6. Bunnel, H. T., Lilley J.: Schwa variants in American English. In: Proceeding of the InterSpeech 2008. Brisbane, Australia (2008) 1159−62
7. Cruttenden, A.: Intonation (2nd ed.). Cambridge University Press, New York (1997)
8. Girosi F.,: An equivalence between sparse approximation and support vector machines. Neural Computation Vol. 10 No. 6 (1998) 1455–1480
9. Dressler, W. U.: Explaining Natural Phonology. In: Phonological Yearbook 1 (1984) 29−50
10. Flemming, E.: The phonetics of schwa vowels. Manuscript, MIT (2007)
11. Flemming, E., Johnson S.: Rosa's roses: reduced vowels in American English. In: Journal of the International Phonetic Association Vol. 37 (2007) 83−96
12. Gósy M.: Virtuális mondatok a spontán beszédben. In: Gósy M. (szerk.): Beszédkutatás 2003. MTA Nyelvtudományi Intézet, Budapest (2003) 19−44
13. Gósy M.: Fonetika, a beszéd tudománya. Osiris Kiadó, Budapest (2004)
14. Gósy M.: The manifold function of schwa. Grazer Linguistische Studien 62 (2004) 15−26

15. Gósy M.: Magyar spontánbeszéd-adatbázis – BEA. In: Gósy M. (szerk.): Beszédkutatás 2008. MTA Nyelvtudományi Intézet, Budapest (2008) 194–207
16. Harmegnies, B., Poch-Olivé D.: A study of style-induced vowel variability: laboratory versus spontaneous speech in Spanish. In: Speech Communication Vol. 11 (1992) 429−37
17. Heuvel, H., Kuijk D., Boves L.: Modeling lexical stress in continuous speech recognition for Dutch. In: Speech Communication Vol. 40 (2003) 335–50
18. Jason, L.: Data-driven investigation of subphonemic variation: "Front" schwa vs. "back" schwa. Paper read at the Cognitive Science Graduate Student Conference 2008. Delawer, Friday April 18th (2008)
19. Jenkin, K. L., Scordilis M. S.: Development and comparison of three syllable stress classifiers. In: International Symposium on Chinese Spoken Language, ICSLP (1996) 733−6
20. Kálmán, L., Nádasdy Á.: A hangsúly. In: Ferenc Kiefer (szerk.): Strukturális magyar nyelvtan 2: Fonológia. Akadémiai Kiadó, Budapest (1994) 393−467.
21. Kopecký, J., Glembek O., Karafiat M.: Advances in acoustic modeling for the recognition of Czech. Paper read at the International Conference on Text, Speech and Dialogue; TSD (2008)
22. Kohle, K. J.: Prosodic boundary signals in. German. Phonetica Vol. 40 (1983) 89–134
23. Kuijk, D., Boves L.: Acoustic characteristics of lexical stress in continuous telephone speech. Speech Communication Vol.27 No.2 (1999) 95−111
24. Ladefoged P.: A course in phonetics. Third edition. Harcourt Brace Jovanovich, New York. (1993)
25. Ladefoged P., Maddieson I.: Vowels of the world's languages. In: Journal of Phonetics Vol. 18 (1990) 93−122
26. Madelska, L., Dressler W. U.: Postlexical stress processes and their segmental consequences illustrated in Polish and Czech. In: Hurch, B., Rhodes, R. A. (szerk.): Natural Phonology: The state of the art. Mouton de Gruyter, Berlin & New York. (1996) 189−200
27. OSU SVMs Toolbox for MATLAB (http://www.ece.osu.edu/~maj/osu_svm/)
28. Padgett, J., Tabain M.: Adaptive Dispersion Theory and phonological vowel reduction in Russian. In: Phonetica Vol. 62 (2005) 14−54
29. Pennington, M. C.: Phonology in English language teaching: An international approach. Longman, London (1996)
30. Szaszák Gy.: A szupraszegmentális jellemzők szerepe és felhasználása a gépi beszédfelismerésben. Ph.D. értekezés, Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék (2009)
31. Simon O.: Anyanyelvi és idegen nyelvi percepciós működések összefüggései az általános iskolában. In: Navracsics J., Tóth Sz. (szerk.): Nyelvészet és interdiszciplinaritás. Generalia, Veszprém (2004) 438−449
32. Siptár, P., Törkenczy M.: The phonology of Hungarian. Oxford University Press, Oxford (2000)
33. Swerts, M., Kloots H., Gillis S., Schutter, G.: Vowel reduction in spontaneous spoken Dutch. In: Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition. Tokyo, Japan (2007) 31–34
34. Valyon J., Horváth G.: Least squares szupport vektor gépek adatbányászati alkalmazása. Híradástechnika Vol. 60 No.10 (2005) 33–38
35. Young, S., Evermann, G., Gales, M.,Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.3). Cambridge University Engineering Department, Cambridge (2005)