Folyamatos beszéd szószintű automatikus szegmentálása szupraszegmentális jegyek alapján

Szaszák György[1], Vicsi Klára[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Kutatólaboratórium
{vicsi, szaszak}@tmit.bme.hu
http://alpha.tmit.bme.hu/speech/

Kivonat: Cikkünkben a folyamatos beszéd szupraszegmentális jegyeken alapuló, szószintű szegmentálási lehetőségeit vizsgáljuk statisztikai megközelítésben, rejtett Markov modellek használatával. A szószintű szegmentálás a folyamatos gépi beszédfelismerés robosztusságát növelheti zajos körülmények között, illetve csökkentheti a keresési teret a dekódolás folyamán. Rendszerünk az alapfrekvencia és az energiaszint értékeit veszi figyelembe, az időtartamok pontos mérése ugyanis felismerési feladatban nehezen kivitelezhető. A rendszert kötött hangsúlyú nyelvekre dolgoztuk ki, és a magyar mellett finn nyelvre is adaptáltuk, illetve vizsgáltuk kétnyelvű rendszerek teljesítményét is, amely a működés hatékonyságát növelte. A statisztikai alapú szegmentáló eredményeit összehasonlítottuk korábbi, szabálybázisú eredményeinkkel, a magyar, illetve a finn nyelv szegmentálási lehetőségeit számos paraméter függvényében vizsgáltuk. Megállapíthatjuk, hogy kísérleteink alapján a kötött hangsúlyú nyelvek esetén a beszéd szószintű tagolása megbízhatóan megvalósítható, ami biztató kilátásokat jelent a kidolgozott rendszer beszédfelismerőbe integrálására vonatkozóan.

1 Bevezetés

A prozódia vagy más néven a szupraszegmentális hangszerkezet az emberi beszéd szerves részét képezi, funkciói részben univerzálisak, részben nyelvspecifikusak. Az univerzális funkciók közül kiemelendő a beszéd értelmezésének megkönnyítését célzó szintaktikai tagolás és a modalitás, de ide tartozik a beszélő érzelmeinek, szándékainak kifejezése is [2]. Ezen univerzális funkciók nyelvenkénti konkrét realizációja már többnyire nyelvspecifikus, míg az “eszközök” sokszor univerzálisak: intonáció, hangsúly, szünetek, ritmus, stb. A szupraszegmentális hangszerkezet segítségével valósíthatja meg a beszélő mondandójának kommunikációs szándékának megfelelő strukturálását. Így, ha a beszéd prozódiai tagolása a szintaxis követelményeinek megfelelően alakul, akkor az egyes szakaszokat prozódiai frázisoknak nevezhetjük. Triviális prozódiai frázis például a két levegővétel közötti beszédszakasz. A műszaki gyakorlatban a prozódiai jegyek reprezentálása három, akusztikailag jól mérhető jellemző révén történhet, ezek az alapfrekvencia, az intenzitás és az időtartam. A szupraszegmentális hangszerkezet egyes elemei – a prozódiai jegyek – lényegében e három akusztikai jellemző különböző

időtartományokra érvényes – értsd szó- vagy mondatszintű – kombinációiként is felfoghatók.

1.1 Rövid történeti áttekintés

A prozódiai jegyek felhasználása beszédfelismerési feladatokban a robosztusság növelésére napjainkban ismét reneszánszát éli. A nyolcvanas évek közepének első próbálkozásai [1,7] során a technikai szint még nem volt adott ahhoz, hogy a kapott eredmények alapján azok a beszédfelismerőkbe is beépíthetőek legyenek. Ez a látszólagos kudarc – Philippe Langlais értelmezésében [3] – elsősorban az alábbi három nehézség miatt következett be:
– a prozódiai tudás jelentős mértékű variáltsága (a beszédtípus, a beszélőtől, a tartalom, a környezet, stb. függvényében);
– a szupraszegmentális szinten hordozott információ és az üzenet nyelvi szerveződési szintjei közötti kapcsolatok bonyolutsága;
– és a prozódiai paraméterek mérésének nehézségei, illetve rendszerbe illesztésük a percepció szintjén.

Néhány korábbi munkában [8,10] a kutatók a prozódiai időtartamok mérésével próbálták meg a beszédbeli határokat detektálni, esetenként a rendszert [8] zajos környezetben működő HMM beszédfelismerő front-end moduljaként megvalósítva. Történtek kutatások [4,5] több prozódiai jellemző alapján készített, folyamatos beszédfelismerő kiegészítő moduljaként működő frázisszintű szegmentálóra is.

1.2 Prozódiai jellemzők a magyar és finn nyelvekben

A beszédképzés során az artikulációs szervek folymatosan mozgásban vannak, amely által folytonos akusztikai jelet hoznak létre. Az ember a beszédértemezés során a szintaktikai és a fonológiai szabályok alapján képes a nyelvi egységek, így a prozódiai frázisok tagolására, a mondatok, szavak azonosítására. Kísérletünkben azt vizsgáltuk, lehetséges-e a szóhatárok nagy hatékonyságú detektálása prozódiai jegyek alapján a folymatos magyar és finn beszédben. Ennek során nagyban kihasználtuk, hogy a mind a magyar, mind a vele rokon finn nyelv kötött hangsúlyú [9]. Emiatt mindig a mondatot felépítő szintagmák első szótagjai kapják a hangsúlyt, amelynek detektálásával kellő pontossággal megtalálhatóak a szószerkezetek, sőt – a kötőszavak, névelők és egyéb hangsúlytalan elemek kivételével –az azokat felépítő szavak határai is. Ilyenformán műszaki szemszögből a tényleges prozódiai frázist szűkebben is értelmezhetjük, amely esetenként egészen az egyes szavak határaiig lebontható alkotó elemekre, a nyelv rétegződésének megfelelően – amelyet kísérleteink tanúsága szerint a magyar nyelv sok esetben még a prozódiával is érzékeltet. A fentiek miatt tartottuk célszerűnek a “szóhatár” kifejezés használatát cikkünkben a “prozódiai frázis” kifejezés helyett.

Kísérleteink során a szabály alapú megközelítés esetéhez hasonlóan [9] ismét az alapfrekvencia és az energiaszint feldolgozását tartottuk célszerűnek és kivitelezhetőnek. Az 1. ábra magyar nyelvű példamondatán jól követhető, ahogyan az alapfrekvencia és az energiaszint a szóeleji hangsúlyoknak megfelelően jelentősen emelkedik a szóhatárok után. A szótagok magánhagzóinak hossza nem mutat

egyértelmű szabályszerűséget. Ebben az esetben az alapfrekvenciát, az energiaszintet és az időtartamokat a szótagok magánhangzóinak közepén mértük.

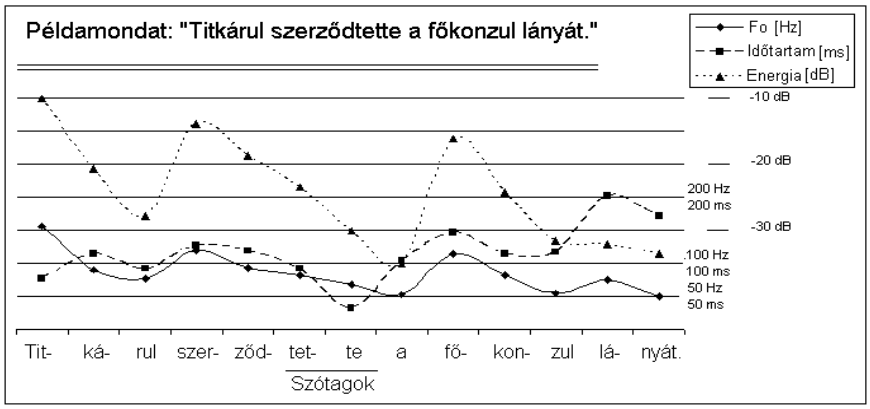

Fig. 1. A magánhagzók közepén mért alapfrekvencia, energiaszint és időtartamok szótagonként a "Titkárul szerződtette a főkonzul lányát" magyar mondatban.

2 Vizsgálati metodika

A szóhatár detektálást a bevezetőben leírtak értelmében prozódiai szegmentálásra vezettük vissza. A korábban elvégzett szabálybázisú vizsgálatok során a hangsúlyt detektáltuk [9], míg a statisztikai alapú automatikus szóhatár meghatározás esetére a HTK [11] rejtett Markov modellek generálására és tesztelésére kialakított fejlesztőkörnyezetet használtuk. Bár a HTK eredetileg beszédfelismerési célokra készült – tehát a fonémák akusztikai HMM modelljeinek elkészítésére és tesztelésére összpontosít –, a benne foglalt HMM implementáció mégis hatékonyan használható más típusú, rejtett Markov modellel leírható osztályozási feladatokra is. Mindezt a laboratóriumunkban kifejlesztett, a HTK programmodul elé illesztett előfeldolgozó egység biztosítja, amely a HTK által értelmezhető formátumúvá konvertálja az adatokat, helyettesítve a beszédfeldolgozás lényegkiemelő modulját.

Vizsgálataink alapjául a BABEL [6], magyar nyelvű beszédadatbázist, és a Helsinki Műszaki Egyetem finn nyelvű beszédadatbázisát (FSD) [12] használtuk fel. Mindkét adatbázis – a későbbiekben részletezendő – prozódiai szintű szegmentálását szakértő végezte. Az adatbázisok magyar nyelvre 22 beszélőtől 1600 mondatot, finn nyelvre 4 beszélőtől 250 mondat tartalmaztak.

2.1 Akusztikai előfeldolgozás

A felhasznált prozódiai jellemzők az alapfrekvencia (Hz) és az energiaszint (dB). Az alapfrekvencia számításakor az autokorrelációs módszert használtuk: az x(n) diszkrét jel autokorrelációs függvénye:

$$R(k) = \sum_{k=N-n-1}^{N} x(n)x(n+k) \quad (1)$$

Az F_0 alapfrekvencia értékét az *i*-edik keretre medián szűrés kapjuk az alábbiak szerint (a keretképzési idő 25,6 ms volt):

$$F_0(i) = \textbf{med}\{ F_0(i-3), F_0(i-2), F_0(i-1), F_0(i), F_0(i+1), F_0(i+2), F_0(i+3) \} \quad (2)$$

Az E(i) energiaszint számítása 100 ms integrálási idővel történt:

$$E(i) = \frac{1}{M} \sum_{n=i-\frac{M}{2}}^{i+\frac{M}{2}} x^2(n) \quad (3)$$

ahol M a 100 ms-ra eső minták száma. Az energiaszint értékek keretideje szintén 25,6 ms.

Az így kiszámított alapfrekvencia- és energiaszint-értékeket használjuk a betanításhoz, az első és második deriváltak hozzáfűzése után, illetve a prozódiai alapon működő szó szintű szegmentáló is ilyen bemenő adatokat vár.

2.2 A statisztikai megközelítés

A statisztikai megközelítés során az egyes Markov modellek meghatározott intonációs osztályokra készülnek, amely esetünkben megadja az adott szószerkezet dallamsémáját.

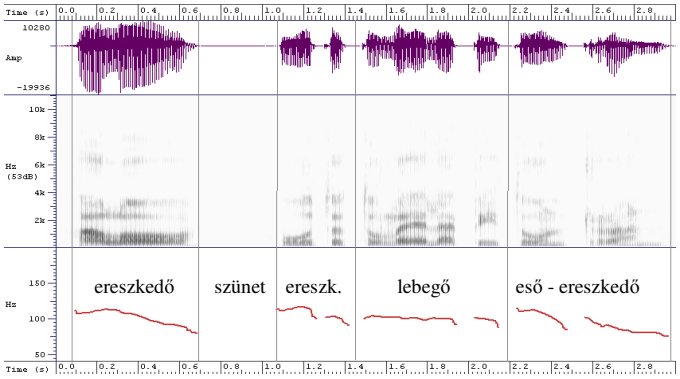

Fig. 2. Néhány tipikus intonációs szegmenstípus magyar nyelvre. Az időfüggvény (felső sáv) és a spekrum (középső sáv) mellett az alapfrekvenciát az alsó sávban láthatjuk.

Kísérleteink során úgy találtuk, hogy célszerű a megkülönböztetett intonációtípusok számát alacsonyan tartani, így csupán az 5, hagyományosnak

nevezhető osztályt különböztettünk meg, amelyek az ereszkedő, az emelkedő, az eső, a szökő és a lebegő. Ehhez hozzávéve a szünetet a kapott 6 féle HMM modellt használtuk vizsgálatainkban. A betanító anyag prozódiai szintű szegmentálásakor ezt a 6 típust jelölték a szakértők ügyelve arra, hogy minden szegmenshatár szóhatárra kerüljön. A 2. ábrán látható példaként néhány tipikus intonációs típus. A szegmentálás során tulajdonképpen az intonációs frázisok határainak bejelölése történt.

A betanítás, illetve az automatikus szószintű szegmentálás blokksémája a 3. ábrán látható. A működés a HMM beszédfelismerő rendszerrel analóg azzal a különbséggel, hogy mások a bemenő adatok, emiatt más az előfeldolgozás, a kimenetből pedig csak a szóhatárok időbeli elhelyezkedése releváns.

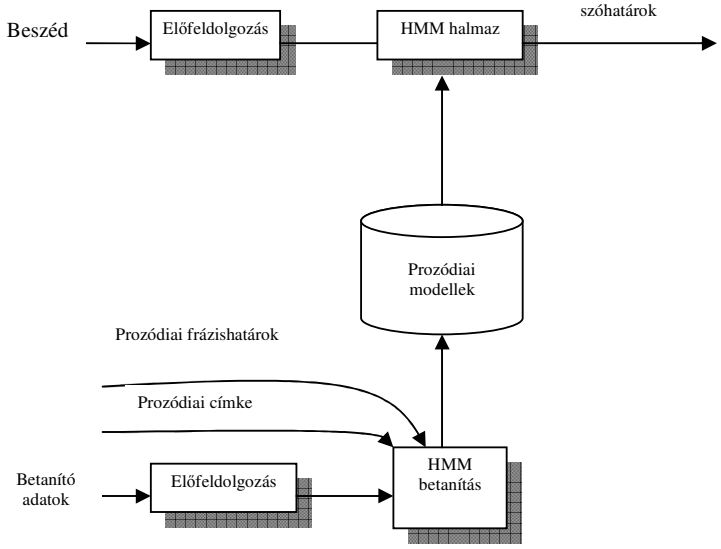

Fig. 3. Az automatikus szószintű szegmentáló vázlatos felépítése, a betanítás és a működés blokksémája

2.3 Kiértékelés

Az eredmények kiértékelésénél – a szabálybázisú megközelítéshez hasonlóan [9] – két mutatót használunk. A **pontosság** azt adja meg, hogy a szupraszegmentális jegyek alapján automatikusan detektált szóhatár az esetek hány százalékában valóban szóhatár. Ez a mutató a beszédfelismerésnél elterjedten használt WER mutató inverz megfelelője a prózódia esetére. A másik mutató, a **hatékonyság** pedig az összes szóhatárok közül a megtaláltak arányát adja meg százalékosan.

Nyilvánvaló, hogy a hatékonyságra jóval 100% alatti értékeket fogunk kapni, hiszen az egy hangsúlyozási-hanglejtési szakaszban lévő szókapcsolatokat sokszor nem tudjuk a prozódia alapján elkülöníteni. Szintén nem lesz lehetséges a névelők, a rövidebb kötőszavak, stb. pontos elkülönítése.

A fontosabb mutató a pontosság. Nyilvánvaló, hogy a hibás szóhatár detekció rontaná a csatlakoztatott beszédfelismerő teljesítményét, ezért ennek az értéknek a maximalizálása kritikus. A szabály alapú megközelítés esetén a pontosság növelése a hatékonyság csökkenését vonta maga után [9], azonban a pontosság értékét ezen az áron is – ésszerű keretek között – maximalizálni kellett.

A szóhatár detekciót akkor tekintettük helyesnek, ha az a valódi szóhatár 100 ms-os környezetébe esett. Az összehasonlítás a betanító anyagban nem szereplő anyag alapján történt, a szószintű szegmentáló kimenetét viszonyítottuk a tényleges szóhatárokhoz.

3. Eredmények

Az eredmények ismertetése mellett szeretnénk kitérni a rendszer főbb paramétereinek optimalizálási lépéseire is, a kapott eredményeket pedig összehasonlítjuk a magyar nyelvre kapott szabálybázisú hangsúly-detektálás esetében kapottakkal, illetve értékeljük a magyar és a finn nyelvre adódott eredményeket.

3.1 A HMM modell struktúrájának optimalizálása

Az alpontban a HMM intonációs modellek két fontos jellemzőjének, az állapotok számának és a kibocsátási valószínűség eloszlást leíró Gauss függvények számának optimalizálási lépését mutatjuk be röviden.

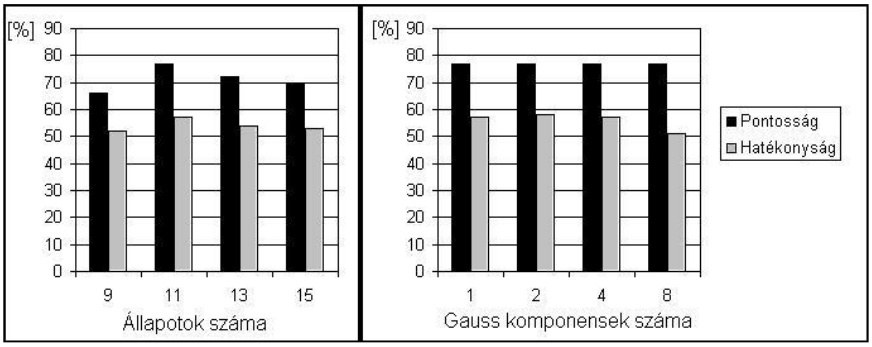

Fig. 4. A pontosság és hatékonyság alakulása az állapotszám függvényében (balra, 4 férfi beszélő, 2 Gauss), illetve a Gauss komponensek számának függvényében (jobbra, 4 férfi, 11 állapot).

A modellek állapotszáma 9 és 15 között változhat, amelyből az első és az utolsó nem kibocsátó állapotok, átmenet mindig csak a következő állapotba lehetséges. Célszerű, ha a modellek legalább 9 állapotúak, hiszen az intonációs frázisok a fonémáknál – amelyeket hagyományosan 5 állapotú modellekkel írnak le 10 ms keretképzési idő mellett – jóval hosszabbak. A 9 állapot kb. 230 ms intonációs frázishossznak felel meg. Esetünkben figyelembe véve a 25,6 ms keretidőt az állapotok számára vonatkozóan ésszerű felső korlát 15 állapot körül van, hiszen

minden állápothoz legalább egy keretet hozzá kell rendelnünk. A 15 állapotnak megfelelő intonációs frázis minimálisan szükséges hossza 380 ms, amely tapasztalataink szerint reális érték. Az optimális állapotszám 11-re adódott (lásd 4. ábra), így a későbbiekben ismertetendő eredményeket is 11 állapotú modellekkel kaptuk. Az egyes állapotokban a kibocsátási eloszlásokat Gauss függvények súlyozott összegével írjuk le [11], a normál függvény komponensek száma esetünkben 1 és 8 között változtatható, tekintettel azonban arra, hogy az alapfrekvencia és az energiaszint menete a beszéd spektrumánál lényegesen egyszerűbb jellemző, elegendőnek bizonyult 1, esetleg 2 Gauss komponens használata (lásd 4. ábra).

3.2 Statisztikai alapú szóhatár detektálás magyar nyelvre

A statisztikai alapú szóhatár meghatározás esetére két betanítási stratégiával kísérleteztünk. Az első esetben vagy csak az alapfrekvencia, vagy csak az energiaszint adataival dolgozott a rendszer, az első és a második deriváltak kiszámítása után csak az egyik prozódiai jellemző (3 elemű jellemzővektor) alapján történt szóhatár detekció. A második esetben mind az alapfrekvencia, mind az energiaszint értékei, első és második deriváltjai alapján történt a betanítás (6 elemű jellemzővektor). Az eredmények a várakozásoknak megfelelően ez utóbbi esetben jobbak, amint azt az 1. táblázatban össze is foglaltuk. A betanítás 14 magyar férfi beszélő anyagával, míg a tesztelés 18 magyar férfi beszélő anyagával történt. A pontosság akkor nagyobb, ha mind az alapfrekvencia, mind az energiaszint típusú értékeket figyelembe vesszük, igaz így a hatékonyság 5-10%-kal csökken.

1. táblázat: statisztikai alapú, automatikus szóhatár detektálás pontossága és hatékonysága magyar nyelvre a bemeneti paraméterek függvényében, 11 állapotú, a kibocsátási valószínűséget 1 Gauss függvénnyel leíró rejtett Markov modellekkel

Prozódiai jellemzők	Nyelv	Betanító anyag	Teszt-anyag	Pontosság [%] / hatékonyság [%] (11 állapotú modell, 1 Gauss)
$F_0+\Delta F_0+\Delta^2 F_0$	magyar	14 férfi	18 férfi	67.4 / 58.4
$E+\Delta E+\Delta^2 E$				67.4 / 63.9
$F_0+\Delta F_0+\Delta^2 F_0$ $+E+\Delta E+\Delta^2 E$				**76.5 / 53.0**

2. táblázat: statisztikai alapú, automatikus szóhatár detektálás pontossága és hatékonysága magyar nyelvre a betanító anyag méretének függvényében, 11 állapotú, a kibocsátási valószínűséget 1 Gauss függvénnyel leíró rejtett Markov modellekkel

Prozódiai jellemzők	Nyelv	Betanító anyag	Teszt-anyag	Pontosság [%] / hatékonyság [%] (11 állapotú modell, 1 Gauss)
$F_0+\Delta F_0+\Delta^2 F_0$ $+E+\Delta E+\Delta^2 E$	magyar	1 férfi	18 férfi	77.3 / 46.4
		4 férfi		**77.4 / 57.1**
		14 férfi		76.5 / 53.0

Megvizsgáltuk azt is, mekkora betanító adatbázissal lehet a legoptimálisabb eredményt elérni. A betanító anyagot így először 4, majd egyetlen férfi beszélőre szűkítettük, és ugyanazon feltételekkel, ugyanazon 18 férfi beszélő anyagával tesztelést végeztünk. A betanító anyagot ebben az esetben gondosan választottuk ki, különösen ügyelve arra, hogy a betanításhoz használt beszédminták kellően tagoltan, helyes hangsúlyozással beszélő személytől származzanak. Az eredményeket a 2. táblázatban foglaltuk össze. Meglepő, hogy a pontosság gyakorlatilag függetlennek tekinthető a betanító anyagban szereplő beszélők számától, ugyanakkor a hatékonyság már függ ettől, optimálisnak a 4 férfi beszélő anyagával végzett betanítás adódott, ekkor 77,4% pontosságot értünk el 57,1% hatékonyság mellett. Ezek az eredmények felülmúlják a szabálybázisú megközelítéssel kapott értéket, amely esetében 77% pontosság mellett a hatékonyság csupán 23% volt. (vö. [9]). A statisztikai alapú szószintű szegmentáló kimenetére példát a 4. ábrán mutatunk be.

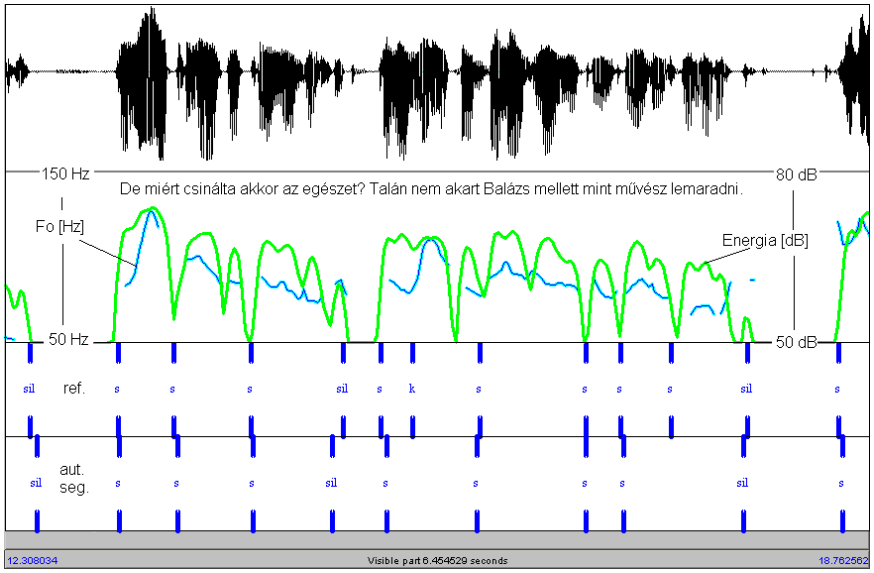

Fig. 4. A prozódiai szegmentálás (3. sáv) és az automatikus szegmentálás (4. sáv) összehasonlítása a „De miért csinálta akkor az egészet? Talán nem akart Balázs mögött mint művész lemaradni.“ szövegrészleten. Az ábra felső részén az időfüggvény (1. sáv), valamint az alapfrekvencia és az energiaszint menete (2. sáv) is látható.

3.3 Statisztikai alapú szóhatár detektálás finn nyelvre

Finn nyelvre a magyar nyelv esetében bemutatott eljárást követve végeztük mind a betanítást, mind a tesztelést. Erre az esetre is a 11 állaptú, 1 Gaussos HMM modellek adták a legjobb eredményt mind az alapfrekvencia, mind az energiaszint, valamint

ezek első és második deriváltjai alapján. A kapott eredményeket a 3. táblázatban foglaltuk össze.

A 3. táblázatból látható, hogy finn nyelv esetén a pontosság alacsonyabb, 69.2%, ugyanakkor a hatékonyság jóval nagyobb, 76.8%, mint a magyar nyelv esetében. Ennek magyarázata az lehet, hogy a kísérleteinkben felhasznált finn beszédet a magyarnál lényegesen lassabb beszédtempó jellemzi, illetve rendkívül gyakoriak a finnben a hosszú, felpattanó zárhangok. Ezeken a helyeken az alapfrekvencia és az energia is hasonlóan viselkedik, mint a szóhatárokon, így az alacsonyabb pontosság abból adódik, hogy a szavak belsejében a hosszú felpattanó zárhangokat is szóhatárként detektálja a rendszer. Mindezt a szegmentáló kimenete is visszaigazolja, hiszen a tévesen detektált szóhatárok finn nyelv eseténben gyakran a hosszan ejtett felpattanó zárhangok zár szakaszára estek. A nagyobb hatékonyság ugyanennek a következménye: a lassúbb beszédtempó miatt a szóhatároknál jobban érzékelhető a szünet, illetve az alapfrekvencia és az energiaszint leesése, így jóval több szóhatárt találunk meg. Véleményünk szerint finn nyelv esetén gyakorlatilag csupán a névelők, kötőszavak előtt, és az egybeolvadásra hajlamos jelzős szerkezetek között nem detektálja a rendszer a szóhatárt, ennek alátámasztása azonban további ellenőrzést igényel.

3. táblázat: statisztikai alapú, automatikus szóhatár detektálás pontossága és hatékonysága finn nyelvre, 11 állapotú, a kibocsátási valószínűséget 1 Gauss függvénnyel leíró rejtett Markov modellekkel. Az összehasonlításhoz a magyar nyelvű eredményeket is feltüntettük.

Prozódiai jellemzők	Nyelv	Betanító anyag	Teszt-anyag	Pontosság [%] / hatékonyság [%] (11 állapotú modell, 1 Gauss)
$F_0+\Delta F_0+\Delta^2 F_0$ +E+ΔE+Δ^2E	finn	4 fő	4 fő	**69.2 / 76.8**
	magyar	4 fő	4 fő	77.3 / 57.1

3.4 Statisztikai alapú szóhatár detektálás kétnyelvű rendszerrel

A módszer más, kötött hangsúlyú nyelvekre való alkalmazhatóságának próbájára magyar anyagon tanított modellekkel finn beszédet szegmentáltunk, illetve ellenkező irányban is végeztünk vizsgálatokat. Megvizsgáltuk továbbá, hogy milyen teljesítményű a mind magyar, mind finn anyaggal vegyesen tanított kétnyelvű rendszer. Az eredményeket a 4. táblázatban mutatjuk be.

A 4. táblázat eredményeiből az tűnik ki, hogy a magyar anyagon tanított, finn nyelvre használt szegmentáló pontossága megegyezik a finn nyelven tanított és finn nyelven tesztelt rendszer pontosságával, a hatékonyság viszont leromlott. A finn anyagon tanított, magyar nyelvre használt szegmentálók esetében a pontosság leromlik, a hatékonyság nem javul. Ezzel szemben a mindkét nyelvű anyaggal vegyesen betanított rendszer pontossága ugyan nem javul az egynyelvű esetekhez képest – magyarra 75%, finnre 69% –, ugyanakkor a hatékonyság jelentősen nagyobb az egynyelvű esethez képest, magyarnál 57% helyett 68%, finn esetében 76% helyett 83%, ami magyar nyelv esetén 19%-os, finn nyelv esetén 9%-os, tehát igen jelentős hatékonyságbeli javulást jelent.

4. táblázat: statisztikai alapú, automatikus szóhatár detektálás pontossága és hatékonysága finn és magyar nyelvre, kétnyelvű rendszerrel.

Prozódiai jellemzők	Betanító anyag	Tesztanyag (4 fő)	Pont. [%] / hat. [%] (11 állapot, 1 Gauss)
$F_0+\Delta F_0+\Delta^2 F_0$ $+E+\Delta E+\Delta^2 E$	magyar (4 fő)	magyar	77 / 57
	magyar (4 fő)	finn	67 / 52
	finn (4 fő)	magyar	70 / 52
	finn (4 fő)	finn	69 / 76
	vegyes (4+4 fő)	magyar	75 / 68
	vegyes (4+4 fő)	finn	69 / 83

4 Összefoglalás

Az alapfrekvencián, és az energiaszinten, mint szupraszegmentális beszédjellemzőkön alapuló automatikus szószintű szegmentálás igen ígéretes eredményeket adott. Ezek alapján statisztikai alapon, az intonációs frázisok rejtett Markov modell segítségével történő leírásával lehetséges a beszédben a szavak határainak megbízható, azaz kellő pontosságú detektálása, jó hatékonysági mutatók mellett. A statisztikai alapú módszer esetén ráadásul kevésbé kényszerülünk kompromisszumot kötni a pontosság és a hatékonyság között, mint a korábbi, szabálybázisú rendszer esetén. Rendszerünket kötött hangsúlyú nyelvekre dolgoztuk ki, és sikerrel adaptáltuk a magyar mellett a finn nyelvre is. A finn és magyar nyelvekre a kétnyelvű rendszer az egynyelvűvel azonos pontosság mellett jóval hatékonyabbnak bizonyult.

A szupraszegmentális beszédjellemzők alapján történő szóhatár detektálás a gépi beszédfelismerők működését javíthatja. Egyrészt hozzájárulhat a felismerés során a keresési tér szűkítéséhez, esetleg lehetőséget adhat a felismerés során futó Viterbi algoritmus szakaszolására. Másrészt zajos körülmények között robosztusabbá teheti a felismerő működését, ez irányban azonban még további vizsgálatok szükségesek. A közeljövőben kísérleteket szeretnénk végezni a szupraszegmentális jegyeken alapuló szóhatár detektálás beszédfelismerő rendszerbe illesztésére vonatkozólag.

Köszönetnyilvánítás

Köszönetünket szeretnénk kifejezni *Péter Attilának*, egyetemünk végzős hallgatójának a kísérletekben való aktív részvételéért, továbbá *Toomas Altosaarnak*, a Helsinki Műszaki Egyetem Akusztikai és Jelfeldolgozási Laboratóriumának vezetőjének hathatós segítségéért, illetve azért, hogy hozzájárult a finn adatbázis használatához.
A kutatást az OTKA T 046487 ELE és az IKTA 00056 pályázatok keretében végeztük.

Bibliográfia

1. Di Cristo: Aspects phonétiques et phonologiques des éléments prosodiques. Modèles linguistiques Tome III (1981) 2:24-83
2. Gósy Mária: Fonetika, a beszéd tudománya. Osiris Kiadó, Budapest (2004) 182-243
3. Langlais, P., Méloni, H.: Integration of a prosodic component in an automatic speech recognition system. 3rd European Conference on Speech Communication and Technology. Berlin (1993) 2007-2010.
4. Mandal, S., Datta, A.K. and Gupta, B.: Word boundary Detection of Continuous Speech Signal for Standard Colloquial Bengali (SCB) Using Suprasegmental Features. FRSM
5. Peters, B.: Multiple cues for phonetic phrase boundaries in German spontaneous speech. Proceedings 15th ICPhS, Barcelona (2003) 1795-1798.
6. Roach, P., Vicsi, K. et al.: BABEL: An Eastern European multi-language database. International Conference on Speech and Language Processing, Philadelphia (1996)
7. Rossi, M.: A model for predicting the prosody of spontaneous speech (PPSS model). Speech Communication (1993) 13:87-107.
8. Salomon, A., Espy-Wilson, C.Y., Deshmukh, O.: Detection of speech landmarks. Use of temporal information. Journal of the Acoustical Society of America (2004) 115:1296-1305.
9. Vicsi Klára, Szaszák György, Borostyán Gábor: Folyamatos beszéd szó- és frázisszintű automatikus szegmentálása szupraszegmentális jegyek alapján. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2004) 319-326
10. Yang, L.: Duration and pauses as phrase and boundary marking indicators in speech. Proceedings 15th ICPhS, Barcelona (2003) 1791-1794.
11. Young, S. et al.: The HTK Book (for version 3.2). Cambridge University, UK (2002)
12. Vainio, M., Altosaar, T., Karjalainen, M., Aulanko, R., Werner, S.: Neural network models for Finnish prosody. Proceedings of ICPhS 1999, San Francisco (1999) 2347-2350.