Folyamatos beszéd szó- és frázisszintű automatikus szegmentálása szupraszegmentális jegyek alapján

Vicsi Klára, Szaszák György, Borostyán Gábor

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

Abstract. Cikkünkben a beszéd alapfrekvencia- és energiaviszonyainak vizsgálatával arra keressük a választ, lehetséges-e ezen prozódiai beszédjellemzők alapján valamilyen módon a folyamatos beszéd gépi tagolása frázisok, illetve szószerkezetek, szavak szintjén. Mindezzel a folyamatos gépi beszédfelismerő működését segíthetnénk a szavak, szószerkezetek határainak detektálásával, ezáltal jelentősen lecsökkentve a beszédfelismeréskor a dekódolás során a keresési teret. Kitérünk az egyes algoritmusokkal elért eredmények bemutatására is. A vizsgálatokat statisztikai módszerekkel végeztük az olvasott szöveget tartalmazó BABEL beszédadatbázison. Várhatóan spontán beszédet tartalmazó szövegben a döntési biztonság az itt bemutatandóhoz képest csökken.

1 Bevezetés

A beszédet az artikulációs szervek folyamatos mozgatásával hozzuk létre. A produktum, vagyis a levegőben terjedő nyomáshullámok eszerint a folyamatos mozgás szerint alakulnak. Vizsgálva a keltett nyomáshullámok fizikai paramétereit, azt tapasztaljuk, hogy ezek a paraméterek is folyamatosan változnak, például a szavak között nem tartunk szünetet. Ha mindig szünetet tartanánk, beszédünk akadozóvá válna. A beszédben szavak, szószerkezetek határait csak egy nyelv megtanulása után vagyunk képesek észlelni, magasabb szintű agyműködés eredményeként. A prozódiai jegyek, az alapfrekvencia, az intonáció és az időtartamarányok segítik a beszéd tagolását. Míg a mondat modalitásának kialakulásában az alapfrekvencia menetének egyértelműen meghatározó szerepe van, addig a hangsúly esetében már nem ilyen egyértelmű a helyzet. A szubjektív hangsúlyérzet kialakulásában mindhárom jellemző részt vesz, egymással szoros összefüggésben. Például az intenzitás emelkedése fiziológiai okokból maga után vonja az alapfrekvencia növekedését is, mivel a megemelkedett szubglottális nyomás a hangszalagokat egyúttal szaporább rezgésre kényszeríti [1]. Legnagyobb eséllyel az a szótag kelt a hallgatóban hangsúlyélményt, melynek mind alapfrekvenciája, mint intezitása kiemelkedő. E kettő közül is alapvetőbb az alapfrekvencia, mert keletkezhet hangsúlyélmény kiemelkedő alapfrekvencia esetén akkor is, ha az intenzitás a környező szótagok intezitásánál valamivel kisebb.

A magyar nyelv kötött hangsúlyú nyelv, a hangsúly az első szótagon szokott lenni. A kötött hangsúlyú nyelvekben a hangsúly mondatszinten értelmezhető jól, a szószerkezetek, illetve a mondat hangsúlyosságát a beszélő szándéka, valamint a mondatszerkesztés szabályai határozzák meg. Ennek megfelelően szakaszhangsúlyról és mondathangsúlyról beszélhetünk.
Korántsem minden szó hangsúlyos tehát, az azonban biztos, hogy ha a mondat valamely szótagja hangsúlyt kap, ez a szótag szinte biztosan szó elején található [3]. Folyamatos gépi beszédfelismerésnél éppen ezért támpontot adhat az, ha a hangsúlyt detektálni tudjuk, mert tagolni tudjuk a beérkezett fonémafolyamot, valamint abban fix pontokat határozhatunk meg, mely által a felismerés hatékonysága is javul. Vélelmezhetjük továbbá, hogy a hangsúlyos pozícióban lévő fonémák felismerése is biztosabb volt (a hangsúlyozás igen sokszor párosul gondosabb artikulációval [2]).
Nem szóltunk még egy fontos prozódiai elemről, a beszédszünetekről. Szünet alatt a prozódiában nem csak a tényleges akusztikai jelkimaradás értendő, hanem másodlagos szünethordózókat is ismerünk. Ezek lehetnek hangzónyújtás, glottális zár, hirtelen hangmagasság-változás, hasonulások elmaradása, illetve ezek kombinációi [2]. Rejtett Markov modelles felismeréskor az akusztikai jelkimaradásra, mely egyben elsődleges szünethordozó, külön modell készíthető, ennek gépi detektálása tehát megoldható. Egyes másodlagos szünethordozók pedig vizsgálhatók a hangsúly detektálásra használt algoritmusokkal. A szünet detektálásának jelentősége ugyan az, mint a hangsúly esetében: utána biztosan új szó kezdődik, ezáltal a felismeréskor kapott fonémafolyam tagolható.

2 Vizsgálati alapelvek

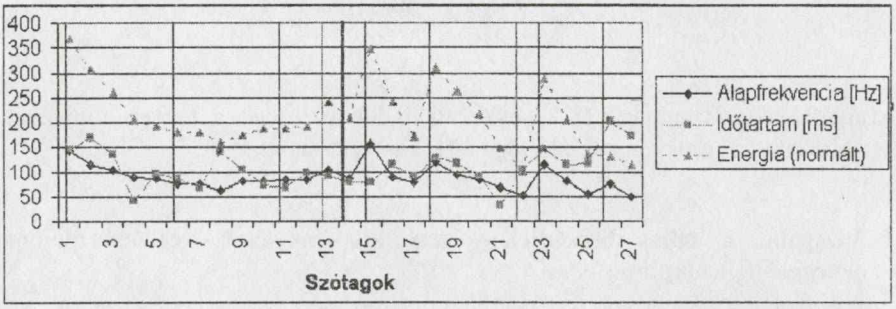
A hangsúly automatikus meghatározásához fizikailag jól mérhető paraméterekre van szükség, előzetes vizsgálataink során az alapfrekvencia és az intenzitásszint értékeit találtuk megbízhatónak. Az időtartam felhasználása két szempontból is problematikus lenne: egyrészt a beszédfelismerés során igen pontosan kellene ismernünk az egyes hangok helyét ahhoz, hogy pontos mérőszámot adhassunk a szótagok hosszára, másrészt a szegmentált anyagok átnézésekor azt tapasztaltuk, hogy a tényleges hangsúllyal a szótagok hossza csak kis mértékben korrelált. A hangsúly végső detektálásához tehát az alapfrekvencia és az energiaszint értékeit figyelembe vesszük (1. ábra).
A hangsúly detektálására kétféle algoritmust dolgoztunk ki. Az első módszernél a hangsúly detektálását csak a szótagok magánhangzóinak kvázistacioner részén mért alapfrekvencia (Hz) és energiaszint (dB) értékeket használjuk. A második módszernél a teljes hanganyagon mért alapfrekvencia és energiaszint értékek alapján történt a hangsúly detektálás. A kétféle algoritmust az alábbiakban mutatjuk be (lásd 2.1. és 2.2. pontokat).

Vizsgálatainkat a BABEL magyar nyelvű, olvasott szöveget tartalmazó adatbázison [4] végeztük, férfi beszélőkre. Az adatbázis fonéma szinten szegmentált, illetve járulékosan szupraszegmentális információkat is tartlmaz. Ezek a szó-, frázis- és mondathatárok helyeinek, valamint a mondattípusok külön jelölését jelentik. A teljes vizsgálat mintegy 1600 mondatnyi anyagon zajlott le, mely 22 beszélőtől származott.

6. ábra: a hangsúly detektálás elvi vázlata

2.1 Hangsúly detektálás szótagok magánhangzóin mért paraméterek alapján

Ebben az esetben a hangsúly vizsgálatakor a szótag magánhangzójának kvázistacioner részét vettük alapul, ezen mértük az alapfrekvencia és az energiaszint értékeit.



7. ábra Alapfrekvencia, energiaszint és időtartam viszonyok a 'Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát.' mondat szótagjainak magánhangzóinak kvázistacioner részein mérve. Az x tengelyen a szótagok sorszáma látható.

A 2. ábrán láthatjuk a 'Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát.' magyar mondat esetén a szótagok magánhanzóinak stacioner részén átlagolt alapfrekvencia és energiaszint-értékeket, valamint a szótag magánhangzóinak hosszát. (Az energiaszint-értékek az egy ábrán való megjeleníthetőség kedvéért lineárisan áttranszformáltak.) A hangsúly detektálásához csúcskeresési algoritmusokat használtunk. Ezek lényege, hogy kiszámítjuk az adott x_i adatsor várható értékét és szórását, majd ezekből egy

$$K = M + k * \sigma \qquad (1)$$

küszöböt határozunk meg, ahol k tetszőleges konstans általában 0.5 – 1.5 közötti értékkel. Ezt követően minden x_i -re megvizsgáljuk, nagyobb-e a K küszöbnél, ha igen, akkor ezt csúcsnak tekintjük, és itt hangsúlyos pozíciót detektálunk.
Magyar nyelvű kijelentő mondatot alapul véve mind az alapfrekvencia-, mind az energiaszint folyamatos csökkenést mutat. Ennek kompenzálására a küszöböt csúszóablakkal számítjuk, az ablak méretét 7 – 17 szótag között célszerű választa-

nunk. Ezáltal a küszöböt a mondat dallammenetéhez igazítjuk. Az *i*-edik szótaghoz tartozó küszöb tehát:

$$K_i = M(x_{i-A}, x_{i-A-1}, ..., x_i) + k * \sigma(x_{i-A}, x_{i-A-1}, ..., x_i) \text{, ha i > A} \quad (2)$$

$$K_i = M(x_1, x_2, ..., x_A) + k * \sigma(x_1, x_2, ..., x_A) \text{ egyébként.} \quad (3)$$

ahol *A* a csúszóablak mérete szótagszámban kifejezve.
Hasonlóan, az egyes szótagok közötti alapfrekvencia- és energia differenciáit is számítottuk, melyeken ugyanezt a csúcskeresést futattuk le azzal a különbséggel, hogy a várható érték (4) és a szórás (5) számításakor a kapott értékek abszolút értékeit vettük:

$$M_i = \frac{1}{A} \sum_{j=i-A}^{i} |\Delta x_j| \quad (4)$$

$$\sigma_i^2 = \frac{1}{A} \sum_{j=i-A}^{i} \left(M_j - |\Delta x_j|\right)^2 \quad (5)$$

A csúszóablakos számítás (2),(3) ekkor is indokolt, mivel a levegő fogytával a beszéddinamika is csökken a frázis vége felé, kijelentő mondatban.

2.2 Vizsgálat a teljes beszédjelen mért alapfrekvencia- és energiamenet jelleggörbéje alapján

Felmerült, hogy a hangsúlydetektáló algoritmust ne csak a szótagok magánhangzóin mért értékek alapján, hanem a teljes hanganyag folytonosnak tekintett alapfrekvencia- és energiaszint-menete alapján próbáljuk megalkotni. Az E_i energiagörbét nagy, 100 ms-os integrálási idővel számítjuk, hogy a gyors, kismértékű fluktuációt kiszűrjük. Ezután ismét átlagoljuk a görbét M = 125 ms-os csúszóablakkal, így kapjuk meg az E_i' görbét (6), majd az eredeti E_i energiagörbe e fölé eső részeit tartjuk csak meg, ebből adódik E_i'' (7).

$$E_i' = \frac{1}{M} \sum_{m=i-\frac{M}{2}}^{i+\frac{M}{2}} E_m, \quad M = 125\text{ms} \quad (6)$$

$$E_i'' = E_i', \text{ ha } E_i \geq E_i' \quad (7)$$

$$E_i'' = 0 \text{ egyébként}$$

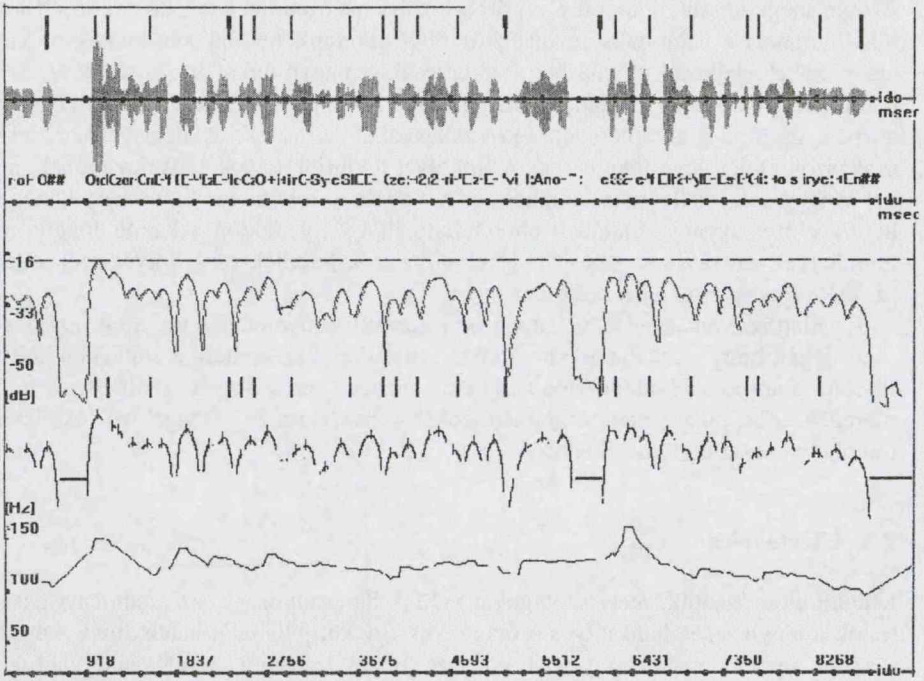
Ezután megkeressük a kapott E'' görbe lokális maximumhelyeit, de ekkor két lokális maximumhelyre minimális távolságküszöböt iktatunk be: ha két lokális maximum ennél közelebb kerül egymáshoz, akkor csak a nagyobbikat fogadjuk el. A lokális maximumhelyek által meghatározott pontokra görbét illesztünk. (3. ábra) A burkológörbén végül negatív meredekségű szakaszokat keresünk. A negatív meredekségű szakaszok elején lévő lokális maximumhelyet tekintjük hangsúlyos pozíciónak, a szóhatárt így ezen lokális maximumhely és a megelőző lokális maximumhely közé jelezhetjük előre, ugyanis általában elmondható, hogy ily módon a kapott lokális maximumhelyek az egyes szótagok magánhangzóinak felelnek meg, mivel ezek energiaszintje a legnagyobb a beszédjelben.

Az alapfrekvencia görbén lényegében hasonló műveleteket végzünk, azzal a különbséggel, hogy a lokális maximumokat közvetlenül az eredeti, a zöngétlen helyeken lineáris interpolációval folytonossá tett görbén keressük. A szóhatár a negatív meredekségű szakasz elején található lokális maximum és a megelőző lokális minimumhely között kerül detektálásra.

2.3 Kiértékelés

Miután előrejeleztük, mely szótagokat vélünk hangsúlyosnak, az eredményt összevetettük a ténylegesen hangsúlyos szótagokkal, így két jellemzőt határoztunk meg. Egyrészt az algoritmus hatékonyságát, hogy az összes szó hány százalékánál találtuk meg a szóhatárt, másrészt, a pontosságot, hogy milyen pontos volt az előrejelzés, azaz volt-e olyan hangsúlytalan szótag, melyet hangsúlyosnak osztályozott a program. A felhasználás szempontjából ez utóbbi a különösen kritikus érték, ugyanis ha felismerés során erre az osztályozásra szeretnénk támaszkodni, nagy pontosságot kell elérni. Sajnos a BABEL adatbázis nem tartalmazza a hangsúlyok címkézését, így minden szó eleji szótagot potenciális hangsúlyos pozíciónak tekintettünk, a hatékonyság emiatt nehezen értelmezhető mutató, hiszen a valós beszédben sem hangsúlyos minden szóindító szótag. Mindazonáltal az egyes módszerek összehasonlítására alkalmas.

A jelleggörbék alapján végzett detektálás esetében fontos látnunk a különbséget az előző módszerhez képest abban a tekintetben, hogy míg az előbb szótagokon mérve vizsgáltuk azok hangsúlyosságát, addig most a teljes jelfolyamon dolgozunk, melyről azután a hangsúlyos pozíciót szótag szintre vissza kell képezni. Akkor tekintettük szóhatár-predikciónkat sikeresnek, ha annak 100 ms-os környezetébe esett a tényleges szóhatár.



8. ábra szóhatár detektálás illusztrációja: a teljes beszédfolyamon mért időfüggvény (fönt) energia és lokális maximum-helyek (középen) és az alapfrekvencia (lent) görbék alapján.

3 Eredmények

A 2.1. pontban bemutatott csúcskereső algoritmusok használatakor különböző *k* konstans és *A* szótagszámban mért ablakszélesség értékek mellett vizsgáltuk a hangsúlyok detektálásának pontosságát és hatékonyságát.

Hatféle kiértékelést készítettünk, ezek az alapfrekvencia, az energiaszint, valamint az alapfrekvencia- és energiamenet alapján együttesen hangsúlyosnak osztályozott pozíciók, valamint az alapfrekvencia-változás, az energiaszint-változás, illetve ezek együttes megléte esetén hangsúlyosnak vélt pozíciók. Az eredmények az 1. táblázatban láthatóak.

Látható, hogy nagyobb, 10 szótag fölötti csúszóablak-szélesség beállítással a pontosság valamelyest növelhető, ez általában maga után vonja a hatékonyság kismértékű csökkenését is. A *k* konstans értékének növelésével – ahogyan az várható – egyértelműen növekszik a pontosság, de a hatékonyság nagyobb mértékben esik az ablakszélesség növelésekor tapasztaltnál.

9. táblázat. Hangsúly detektálás pontossága és hatékonysága magánhangzók kvázistacioner részén mért paramétereinek alapján

A	k	Pontosság/Hatékonyság [% / %]					
		F_0	E	F_0 & E	ΔF_0	ΔE	ΔF_0 & ΔE
7	0.5	49/44	46/30	46/20	76/24	57/21	82/10
7	0.7	50/39	45/27	46/16	77/23	58/19	83/10
7	0.9	51/33	45/24	47/13	78/21	59/17	86/9
7	1.1	52/28	45/21	47/10	79/20	60/15	87/7
9	0.5	49/41	46/29	45/18	76/24	59/21	84/11
9	0.7	50/36	46/26	47/15	77/22	60/19	83/9
9	0.9	52/32	46/23	47/12	78/21	61/17	83/9
9	1.1	52/27	45/20	47/9	79/19	62/15	85/8
13	0.5	51/39	45/27	46/16	77/22	61/19	84/9
13	0.7	52/34	45/23	46/13	78/20	63/18	84/8
13	0.9	52/28	45/20	46/11	79/19	64/16	87/8
13	1.1	54/24	46/18	49/9	79/17	65/14	88/7
17	0.5	51/38	46/26	46/16	78/21	64/19	86/9
17	0.7	53/33	45/22	47/10	78/19	63/17	86/8
17	0.9	54/28	46/20	49/10	79/18	65/15	86/7
17	1.3	56/20	46/15	52/7	81/15	65/11	90/6

Az eredmények mind az energiaszint, mind az energiaszint-változás esetén jóval gyengébbek az alapfrekvenciával kapottaknál. Ennek részben oka lehet, hogy a magánhangzók energiájának szubjektív észlelése függhet a magánhangzótól. A nyitottabb ajkakkal képzett magánhangzók nagyobb energiájúak, emiatt előfordulhat, hogy hangsúlyos, de kerekítettebb ajkakkal képzett magánhangzó energiája kisebb a hangsúlytalan nyílt magánhangzóénál.

A teljes beszédjel alapfrekvencia-, illetve energiamenet jelleggörbe alapján (lásd 2.2. pont) kapott eredményeket a 2. táblázatban láthatjuk. Az eredményeket az 1. táblázat ΔF_0, ΔE, illetve ΔF_0 & ΔE oszlopokban kapott eredményeivel érdemes összevetnünk, mivel a lokális maximumhelyek megkeresése, illetve a jellegörbék tulajdonságainak vizsgálata is delta dimenziójú paramétereken nyugszik. Mindezek alapján elmondhatjuk, hogy az alapfrekvencia alapján a predikció pontatlanabb, igaz valamivel hatékonyabb is. Az energiaszint esetében egyértelműen a második megközelítéssel kapunk jobb eredményt. Ennek oka a fentiekhez hasonlóan valószínűleg az, hogy a zárt magánhangzókat relatíve hangsúlyosabbnak érzékeljük kisebb energiatartalom esetén is, ezzel a módszerrel azonban jobban megfogható ez a jelenség. Az együttes becslés pontosságában és hatékonyság tekintetében szintén felülmúlja az egyszerű csúcskereséssel kapott értékeket. Az összehasonlításkor azonban legyünk óvatosak, mert a két módszer közül a második megítélésekor a mérési pontatlanság nagyobb.

10. táblázat. Hangsúlyos pozíció detektálása a teljes beszédjelen mért jelleggörbék alapján

Pontosság/Hatékonyság [% / %]		
F_0	E	F_0 & E
70/32	69/34	91/14

Összességében elmondhatjuk, hogy bármely módszer felhasználhatóságához megközelítőleg legalább 85%-os, de lehetőleg minél nagyobb pontosság elérése a kívánatos. Ezt az értéket jelen vizsgálatunkban sikerült elérni, 9-14% közötti hatékonyságot kaptunk ebben az esetben a különböző mérési elrendezésekben. Megjegyzendő, hogy a beszédfolyamat során sem minden szó hangsúlyos, ezért a hatékonyságban véleményünk szerint 50%-ot meghaladó eredményt elérni nem is lehetne.

4 Konklúzió

Cikkünkben a folyamatos beszéd automatikus szegmentálásának problematikáját tárgyaltuk. Bemutattuk ennek szerepét a beszédfelismerés során, valamint a hangsúly detektálására kidolgozott két módszert. A kapott eredmények alapján úgy gondoljuk, érdemes a területen tovább vizsgálódni, a közeljövőben tervezzük az alapfrekvencia és energiaszint paraméterek alapján a hangsúly előrejelezhetőségének vizsgálatát statisztikai módszerekkel, valamint a hangsúlydetektáló modult beszédfelismerőbe építve vizsgálni szeretnénk, javítható-e és milyen mértékben a beszédfelismerés hatékonysága.

Köszönetnyilvánítás
A kutatás az OTKA T 046487 ELE és az IKTA 00056 pályázatok keretén belül készül.

Irodalomjegyzék

1. Kassai Ilona: Fonetika. Nemzeti Tankönyvkiadó, Budapest (1998)
2. Kassai Ilona – Fagyal Zsuzsanna: Hogyan észlelik a magyar beszéd szüneteit magyar és francia anyanyelvű hallgatók. In: Magyar nyelvőr, Budapest (1996/120) 209-220. o.
3. Kiefer Ferenc (szerk): Struktúrális magyar nyelvtan, II. kötet, Fonológia. Akadémiai Kiadó, Budapest (2000).
4. Vicsi Klára – Vig Attila: Az első magyar nyelvű beszédadatbázis, Beszédkutatás'98, Tanulmányok az elméleti és alkalmazott fonetika köréből. MTA Nyelvtudományi Intézet, Budapest (1998) 163-178. o.