Kétszintű algoritmus spontán beszéd prozódiaalapú szegmentálására

Beke András[1], Markó Alexandra[2], Szaszák György[3], Váradi Viola[2]

[1] MTA Nyelvtudományi Intézet
[2] ELTE BTK Fonetikai Tanszék
[3] BME Távközlési és Médiainformatikai Tanszék
e-mail:andras.beke@mta.nytud.hu

Kivonat Cikkünkben egy kétlépcsős automatikus szupraszegmentális frázisszegmentálót mutatunk be spontán beszédre. Miután az olvasott beszédre kidolgozott eljárások spontán beszédre nem működnek megfelelően, illetve a felügyelet nélküli kontúrklaszterezés sem hozott kielégítő eredményt, részletesebben is áttekintjük a vonatkozó irodalmat, ami alapján a szünetekkel határolt intonációs frázist tekintjük modellezési alapegységnek. A második szinten beágyazott fonológiai frázisok detektálását vizsgáljuk alapfrekvencia, energiamenet és magánhangzóidőtartamok alapján. A fonológiai frázisok detektálása az akusztikai jellemzők időbeli változását megragadó, szimmetrikus Kullback–Leibler-távolság alapú metrikával képzett különbségjelére meghatározott adaptív küszöbértékkel történik. A fonológiai frázisok detektálási pontossága spontán beszédben 80% körül adódik, percepciós szempontú címkézéssel összevetve. A kétszintű detektálási feladatban az energiaszint és a szünetek az intonációs frázisokban, az alapfrekvencia pedig a fonológiai frázisokban tűnik meghatározónak. Az időtartamok hozzájárulása a frázishatárok észleléséhez – az olvasott beszédhez hasonlóan – spontán beszédben sem igazolható.

Kulcsszavak: prozódia, spontán beszéd, automatikus klaszterezés

1. Bevezetés

A prozódiai frázis és/vagy prozódiai egységek határainak detektálása jelentős és fontos kutatási kérdés. Az elmúlt évtizedek során kifejlesztettek néhány olyan rendszert, amely olvasott, illetve félspontán beszédben képes automatikusan jelölni prozódiai egységek határait [8]. A magyar nyelvű fejlesztésekről az MSZNY konferenciákon is rendszeresen beszámoltunk [26,21]. Spontán beszédben ezek a felügyelt gépi tanuláson alapuló eljárások sokkal kisebb hatékonysággal és pontossággal használhatók [1], emellett jelentős problémaként merül fel a modellezési alapegységnek a megválasztása.

1.1. A prozódiai alapegységről

A beszédprozódia tudományos vizsgálatának elengedhetetlen feltétele annak tisztázása, hogy mit tekintsünk a prozódia egységeinek. A szegmentális fonetikai

megközelítésben vitán felül áll a beszédhang mint alapegység (annak ellenére, hogy a hangátmenet-hanghatár problematikája természeténél fogva megoldatlan marad), a szupraszegmentális megközelítésben azonban az alapegységet illetően sem elnevezésében, sem terjedelmében/tartalmában/definíciójában nincs konszenzus a kutatók között. A spontán beszéd vizsgálatában alapvető problémát jelent a hangfolyam tagolása, annak a közlésegységnek a kijelölése, amely „méreténél" és információtartalmánál fogva alkalmas arra, hogy a beszéd prozódiai alapegységének tekintsük. A legelterjedtebb elképzelés mind a hazai, mind a nemzetközi szakirodalomban a hierarchikus felépítés [19,23,14,12]. Felülről lefelé haladva a következő szintek különíthetők el: megnyilatkozás, intonációs frázis, fonológiai frázis, fonológiai szó, láb, szótag. Felmerül a kérdés, hogy spontán beszédben is alkalmazható-e ez az elkülönítés.

A fentieknek megfelelően jellegzetes különbségeket mutatnak azok a szakirodalmi források, amelyek felolvasott szövegek vagy hipotetikus közlések kapcsán foglalkoznak a tagolással, és azok, amelyek spontán beszéd prozódiai elemzése alapján kísérlik meg megállapítani az egységhatárokat. Az előbbi csoportba tartozó tanulmányok egyértelmű megoldásokat kínálnak. Például Elekfi 1962-es tanulmányában [7] azt mondja, hogy „nagyobb értelmi és kritikai egységeket kell keresni, melyekből a mondat már viszonylag közvetlenül felépíthető. Alkalmas egységnek látszott erre a beszédütem", amely nagyjából a klitikumos egységnek felel meg. Bolla Kálmán a szupraszegmentális alapegység meghatározásakor [3] ezt írja: „A szegmentális szerkezet struktúraképző alapeleme a beszédhang, míg a szupraszegmentális hangszövet legkisebb szerkezeti építőblokkját szupraszegmentális hangszerkezetnek nevezzük", majd megjegyzi, hogy hasonló értelemben használják még az intonációs szerkezet kifejezést is. A bollai hangszerkezet ez alapján tehát analóg a más szerzők által intonációs frázisként tárgyalt alapegységgel. Alapegységnek tehát a szupraszegmentális hangszerkezetet (intonációs frázist) tekinti, de ugyanakkor bizonyos értelemben alapegységként kezeli az ezt hordozó beszédszakaszt is, amelynek mibenlétét, definícióját azonban nem adja meg. A beszédszakasz a fentiek alapján a szintagma és a megnyilatkozás közötti egységet jelentheti, tehát valószínűsíthetjük, hogy a wachai megnyilatkozásegységnek feleltethető meg. Bollánál azonban nem találunk olyan kritériumrendszert, amely alapján meghatározhatnánk az alapegység határait.

Olaszy Gábor rádiós (felolvasott) szövegműfajok – hírek, novella, mese, reklám – komplex akusztikai elemzését végezte el [20]. A legnagyobb szövegegység, amelyen a prozódiát vizsgálta, a mondat volt. Ezeken belül úgynevezett prozódiai egységeket (PrE) határozott meg, amelyeket a szünettartás alapján határozott meg (tehát a mondat eleje-vége, illetve a mondat belsejében tartott szünet jelölte ki a PrE-k határát). A PrE-ket az alapfrekvencia alapján bontotta tovább hangsúlyközi szakaszokra, ezeket intonációs frázisoknak (IF) nevezte.

Élőbeszéddel dolgozott többek között Wacha [27], aki a megnyilatkozást tekinti alapegységnek, s a következőképpen definiálja: „Megnyilatkozáson az élőszóbeli, a (spontán) beszélt nyelvi közlésegésznek (szövegnek, szövegegésznek, beszédműnek) azt a – pontosan többé-kevésbé – elkülöníthető kisebb részét/egységét értem, melyet az írott nyelvhasználatról szólva a mondat, szövegmondat

terminussal szokás megnevezni. A megnyilatkozás a beszélt nyelvnek-nyelvhasználatnak olyan mondat értékű része tehát, melynek határait (kezdetét és végét) utólag – az elhangzó szöveg lejegyzésekor (átírásakor) – állapítottuk meg és jelöltük meg írásjelekkel, figyelembe véve az írásbeliség alapján kialakult mondatfelfogást (konvenciót) is." A megnyilatkozások megnyilatkozásegységekből állnak, így ezek tekintendők alapegységeknek, ugyanakkor Wacha használja a fonemikus frázis kifejezést is „hangképzési szünettől hangképzési szünetig tartó, csöndekkel határolt beszédszakasz" értelemben. Wacha problematikusnak tartja a megnyilatkozáshatárok megállapítását, végül az alábbi öt tényező közül valamely kettőnek az együttes jelenléte esetén tekint befejezettnek egy megnyilatkozást: „1. akusztikus zár, illetőleg ennek hiánya: a közlés dallama úgynevezett pont-hanglejtéssel zárul-e vagy sem, azaz a beszéd dallama mélyre szálló hanglejtéssel jelzi-e a megnyilatkozás befejezését, vagy esetleg nyitva tartott dallammal a folytatást ígéri (a beszédét vagy a megnyilatkozását); [...] 2. grammatikai zár: a megnyilatkozás grammatikailag befejezettnek, esetleg kereknek tekinthető-e vagy sem; [...] 3. követi-e az intonációs vagy grammatikai zárat új megnyilatkozást (új megnyilatkozás kezdetét) jelző intonációs indítás; [...] 4. követi-e vagy sem a megnyilatkozást valamiféle kiegészítő hozzátoldás (pl. értelmező, valamilyen »mondatrész«); [...] 5. a megnyilatkozás-sorozatot megtöri, megszakítja-e szünet vagy sem?" Wacha szerint a megnyilatkozáshatár legbiztosabb jelzője az új megnyilatkozás kezdetét jelző intonáció megjelenése.

Németh T. Enikő [16] a szóbeli diskurzusok megnyilatkozáspéldányokra tagolásakor arra a következtetésre jut, hogy figyelembe kell venni a dallam- és szünetprozodémák mellett a mondattal való kapcsolatba hozhatóságot, valamint pragmatikai szempontokat is. Ez utóbbiak között tekintetbe veszi az interperszonális funkciót, a beszédtettsorozatok összetételét és a pragmatikai kötőszókat, valamint az attitudinális funkciót (és az azt mutató nyelvi eszközöket).

Varga László [24,25] meghatározása fonológiai szempontú, s mint ilyen, a fonológiai szabályrendszer működési hatóköréül szolgáló nyelvi egység mibenlétét határozza meg, de komplex megközelítése figyelembe veszi a spontán beszéd sajátosságait is. Varga szerint a magyar intonációs frázis minimáldefiníciója: olyan szótagsorozat, amely vagy egy a) függelékdallammal, vagy egy b) (előkészítő dallam +) karakterdallammal realizálódik [24]. Ugyanakkor az intonációs frázist olyan egységként is meghatározhatjuk, amely egynél több dallamhangsúlyt fog át, ezt nevezi Varga maximáldefiníciónak, amely szerint az intonációs frázis kétféle lehet: a) függelék típusú, amely csak egy függelékdallamot tartalmaz; illetve (b) hangsúlyos, amely legalább egy karakterdallamból áll (előtte állhat egy vagy több szünet nélküli féleső karakter, s ez(eke)t megelőzheti egy főhangsúlyt nem tartalmazó előkészítő dallam), és szünet követi [25]. Varga tehát egyszerre alkalmazza a minimális (Elekfihez hasonlóan) és a maximális (ahogy Wacha és Németh T. Varga ismeretében teszi) szekvenciákra tagolás elvét, s e kettő közül az utóbbit preferálja, mert az a szintaktikai szerkezetet is tükrözi, ugyanakkor elismeri, hogy bizonyos spontán közlések esetében csak a minimáldefiníció alkalmazható.

Levelt [15] a beszédprodukciós folyamatot középpontba állító munkájában ugyancsak intonációs frázisnak nevezi az alapegységet. Ez megfogalmazása szerint szünettől szünetig tart: a beszélő azzal, hogy (általában több mint 200 ms-os) szünetet tart, befejez egy intonációs frázist. Az intonációs frázis leírásában Levelt öt összetevőt mutat be. A szünettartás bizonyos mértékig a beszélő döntésén múlik, ha jól érthető kíván lenni a hallgató(k) számára, lassan és rövid, akusztikai kulcsokban gazdag intonációs frázisokban beszél. A másik fontos tényező tehát a beszédtempó, amely természetesen befolyásol(hat)ja az intonációs frázisok tartamát is. Az intonációs frázisok megközelítőleg azonos időtartamban valósulnak meg a spontán beszédben (izokrónia), amit az artikulációs program végrehajthatósága indokol (az artikulációs tároló mérete és/vagy a rendelkezésre álló levegő mennyisége). A szintaktikai és a szemantikai szerkezet, valamint végül a beszédtervezés működési feltételei, minősége is nagymértékben hatással vannak az intonációs frázisok egymásra következésére.

Az eddig részleteiben áttekintett elméleti háttérirodalom alapján spontán beszédben is az intonációs frázis kínálkozik egy lehetséges alapelemként a prozódia modellezésében. Jelen munkában az intonációs frázis domináns akusztikai markereként a szünetet jelöljük ki, és kísérletet teszünk az intonációs frázisba mélyebb szinten beágyazott fonológiai frázisok detektálására is, a modellezésben is megtartva ezt a kétszintű hierarchiát. A fonológiai frázisok detektálásában prozódiai-akusztikai jellemzők követésére koncentrálunk eseménydetekciós attítűddel.

1.2. Modellezési megfontolások

A korábbi kísérletekben alkalmazott felügyelt tanulási módszerekkel a gépi tanulást címkézett adatokon végeztük el, amelynek a végén kialakult az osztályozó vagy detektáló, amely képes előre jelezni a prozódiai határokat a prozódiai egységekből származó akusztikai-prozódiai jellemzők alapján. Ez a megközelítés azt is feltételezi, hogy a detektálni kívánt egységek jól definiáltak (határaik és típusaik), és a priori ismertek, amelyek pontosan jelölve vannak a tanító korpuszban.

Újabban jelentős figyelem összpontosul a prozódiai egységek felügyelet nélküli modellezésére [5], vagy a már meglévő, felügyelt tanításból származó prozódiai modellek felügyelet nélküli adaptálására [22]. Ezek a megközelítések akkor is hatékonynak bizonyulhatnak, ha a modellezendő alapegység kérdéses.

Jelen kutatásban arra összpontosítunk, hogy nem felügyelt megközelítésben vizsgáljuk a spontán beszéd prozódiaalapú tagolhatóságát, majd ezt összevetjük intonációs, illetve fonológiai frázisokra való címkézéssel.

2. Anyag és módszer

A kutatáshoz a spontán beszédmintákat a BEA (BEszélt nyelvi Adatbázis [11]) szolgáltatta. 8 beszélő spontán narratíváit választottunk ki, 4 férfi és 4 női beszélőtől. Az így nyert korpuszt 2 fonetikus szakember felcímkézte, ezt a címkézést

kizárólag a kiértékelésnél, referenciaként használjuk. A címkézést 3 szinten végezték: intonációs frázisra (IF), fonológiai frázisra (FF), illetve szószintű átiratra. A vizsgálatokhoz használt korpusz összesen 398 IF-t és 751 FF-t tartalmazott.

2.1. Szegmentálás intonációs frázisokra

A BEA-ban a beszédfordulók szegmentáltak, a beszédfordulók megnyilatkozásokra bontása azonban nem egyértelmű, ahogyan – mint arra korábban már utaltunk – a megnyilatkozás definíciója sem. A beszéd intonációs frázisokra tagolásának irodalmát áttekintve kiolvasható, hogy annyiban viszonylagos konszenzus mutatkozik a kérdéskört vizsgáló kutatásokban, hogy az IF-ok határait, illetve lehetséges határait szünetek, esetenként időtartambeli nyúlás, F0 és energia jelzik. E jellemzők közül a szünetek szerepét több további kutatás is kiemeli [13], magyarra [10] az IF-ok percepciójában.

A fenti megfontolások fényében az IF detektálását energia, spektrális súlypont (centroid) és alapfrekvencia (F0) jellemzők alapján végezzük [9]. A jellemzőket 50 ms ablakkal számítjuk, melyek közül az energia nyilván kisebb lesz szünetekben, a spektrális súlypont pedig robusztusabbá teszi a beszéd/nem-beszéd elválasztását: a magasabb spektrális súlypont általában a beszédszegmensekre jellemző. Az F0-t azért vonjuk be a vizsgálatba, hogy egyrészt robusztusabbá tegyük a beszéd/nem-beszéd detektálását, másrészt lehetőségünk legyen tipikus fráziszáró intonációs markerek követésére is. A jellemzőket normalizáljuk, majd küszöbértéket számolunk, amelynek meghatározására hisztogramanalízist végzünk: a hisztogramokban a leggyakoribb értékeket k-közép klaszterezéssel szeparáljuk, lényegében súlypontokat határozunk meg. Két klaszterközéppontot alapul véve (M_1 és M_2) a küszöbértéket (T) az alábbi összefüggéssel számíthatjuk:

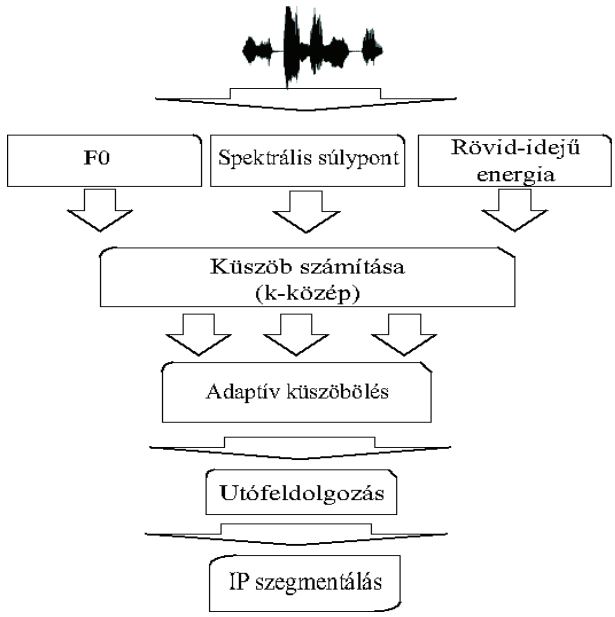
$$T = \frac{W * M_1 + M_2}{W + 1}, \tag{1}$$

ahol W szabadon választható paraméter (értéke kísérleteinkben $W = 0.5$). A küszöbértékkel a jelfolyamon csúcskeresést végzünk, végezetül pedig a csúcskeresés eredményeképpen kapott, keretekre értelmezett detekciós pontokat nagy ablakkal (250 ms) simítjuk, a folyamatot az 1. ábrán illusztráltuk.

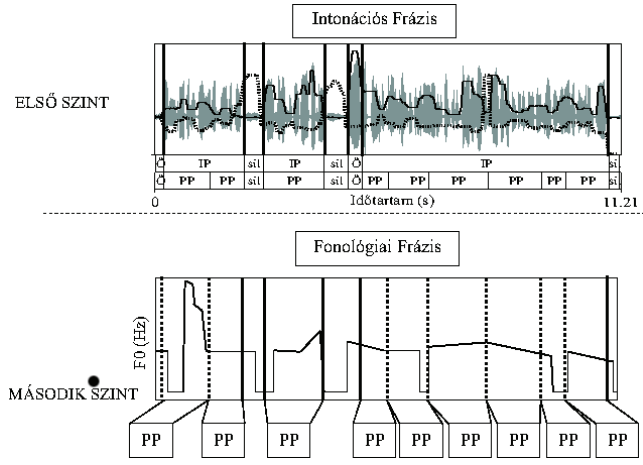
2.2. Szegmentálás fonológiai frázisokra

A FF-ra történő szegmentálás nagyobb kihívás a folyamatos, szünetekkel sem tagolt jelen. Mivel a FF-t az IF-ba beágazottként tekintjük, (2 ábra) a FF-szegmentálás bemenete az IF szegmentálás eredménye. Ily módon a hierarchiát is tükröző eljárást kapunk.

Jellemzőkinyerés. A FF-ok detektálásához a három alapvető prozódiai jellemzőt – alapfrekvencia (F0), energia és időtartam – vizsgáljuk szeparáltan és kombináltan spontán beszédben.



1. ábra. Az IF-szegmentálás vázlata



2. ábra. FF-ok azonosítása a prozódiai hierarchiában

A jellemzőkinyerést az olvasott beszédre is ismertetett módon végezzük [21] F0-ra és energiára, beleértve a részleges extrapolációt is a zöngétlen szakaszokon (kivéve, ha a zöngétlen szakasz hosszabb, mint 150 ms, vagy a zöngétlen szakasz után az F0 értéke eléri a korábbi F0-érték 110%-át).

Az időtartamokat magánhangzók hosszára és magánhangzók közötti távolságra automatikusan számítjuk, e két adatból lényegében a szótaghosszt is megkaphatjuk. Az időtartammérés alapja egy HMM-GMM alapú beszédhang-osztályozó, amely alapvető beszédhang-kategóriákat illeszt a folytonos beszédjelen: magánhangzókat, nazálisokat és approximánsokat, felpattanó zárhangokat, affrikátákat és frikatívákat. A beszédhang-osztályozó front-endje MFCC-jellemzőket számít, első és második deriváltakkal (MFCC39). A back-enden a beszédhang osztálya, kezdő- és végidőpontja jelenik meg.

A kapott magánhangzóhosszakat beszélőnként normalizáljuk, majd időben folytonos (pontosabban 10 ms keretidejű diszkrét) kontúrt állítunk elő simítással, a továbbiakban ezt értjük a tempó alatt.

A beszédhang-osztályozót azért választottuk a tempó kinyerésében, hogy komplett beszédfelismerés végrehajtása nélkül, illetve átíratlan anyagon is működjön a jellemzőkinyerés. Jóllehet maga a beszédhang-osztályozó sem működik hibamentesen, konzekvens hibázása esetén még ki is emeli a fontos jellemzőket – például megnyilatkozás végi irreguláris, kisebb energiájú beszédrészleteken a magánhangzókat nyújtja a többi beszédhang rovására, ez a működési jellegzetesség jelen alkalmazásban még előnyös is lehet.

Az F0, energia és tempó jellemzők mindegyikére első és második deriváltakat is számítunk, majd 0 és 1 közé normalizálunk.

Szegmentálás szimmetrikus Kullback–Leibler-távolsággal. A Kullback–Leibler-távolság (KL) az egyik leggyakrabban használt algoritmus arra, hogy hogyan mérhető a különbözőség két eloszlás között [2]. A KL-távolságot számos területen alkalmazzák mint beszélődetektálás, beszédfelismerés, beszélőfelismerés vagy beszéd/nem beszéd detektálás, stb. Ezek mellett igen népszerű a KL-távolság algoritmus használata szegmentálási problémák megoldására is, mint a beszéd vagy a zene szegmentálásra. A jelen tanulmányban a KL-távolságot a FF-ok határainak meghatározására alkalmazzuk. Matthew és munkatársai [18] kimutatták, hogy a szimmetrikus Kullback–Leibler-távolság egy olyan hatékony távolságmérő eljárás, amely könnyen mérhetővé teszi statisztikailag a különbözőség mértékének kifejezését két beszédjel között. A matematikai hátterét a következőkben ismertetjük: legyen X és Y két random eloszlás, és KL pedig a különbözőség mértéke e két eloszlás között. A távolság $KL(X;Y)$ az X és az Y között a következőképpen számolható:

$$KL(X;Y) = E_X(log(\frac{P_X}{P_Y}), \tag{2}$$

ahol E_X jelöli az X valószínűségi sűrűség függvény várható értékét. Ha az eloszlások Gauss-eloszlással modellezhetők, akkor a fenti egyenlet a következőképpen

alakul:

$$KL(X;Y) = \frac{1}{2}tr[(\Sigma_X - \Sigma_Y)(\Sigma_Y^- 1 - \Sigma_X - 1)] + \\ + \frac{1}{2}tr[(\Sigma_Y^- 1 - \Sigma_X^- 1)(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T], \tag{3}$$

ahol Σ a kovariancia mátrixra, míg a μ a középérték vektorra utal az adott eloszlásban. A Kullback–Leibler-távolság ugyan nemnegatív, de nem valódi metrika, mivel nem szimmetrikus, azaz megkülönböztetheti a modellt és modellezett eloszlást. A KL aszimmetrikus távolságot szimmetrikussá lehet tenni a következő lépéssel:

$$KL2(X;Y) = KL(X;Y) + KL(Y;X). \tag{4}$$

Mint korábban írtuk, ha a két eloszlás Gauss-eloszlással közelíthető, akkor a szimmetrikussá tett formában is létezik $KL2$ szimmetrikus KL távolság.

Jelen munka során a $KL2$-távolságot a beszédjelben egymást követő részek között számoltuk, amely részek 4 keret hosszúságúnak felel meg, vagyis 40 ms időtartamúak. Az ablakhossz 1 keretnyi volt, ami 10 ms-os időtartam. Minden egymást követő beszédrész között számolt $KL2$ érték egy folytonos görbét adott, amely után a következő feladat az volt, hogy ebben a folytonos jelben megtaláljuk a csúcsokat, amelyek a jelen esetben azt jelezték, ahol a két beszédszegmens között a legnagyobb eltérés jelentkezett. A magas $KL2$ érték tehát azt feltételezi, hogy a két beszédszegmens között jelentős az eltérés, míg az alacsony $KL2$ érték az azonosságot feltételezi. A csúcsdetektálás szempontjából igen fontos, hogy milyen ablakhosszban keressük az adott csúcsot. Ezért különböző ablakhosszokat alkalmaztunk, amelyet 10 kerettől (100 ms) 40 keretig (400 ms) növeltük a $KL2$ folytonos görbén. A csúcsdetektálás szempontjából igen fontos feladat a küszöbérték megválasztása is, mivel ettől függ, hogy az adott $KL2$ értéket váltási pontnak fogadjuk el, vagy sem. Ennek megválasztására két adaptív küszöbölési technikát alkalmaztunk (thr_A and thr_B). Az első adaptív küszöbértéket úgy számoljuk, hogy az adott keretben található érték középértékét vesszük, majd megszorozzuk egy konstanssal:

$$thr_A = \alpha \frac{1}{2N_1} \sum(F), \tag{5}$$

ahol F a jellemzővektor, N_1 az ablak hossza, és α a konstans.

Annak érdekében azonban, hogy FF határait detektáljuk, az adott értéknek nagyobbnak kell lennie, mint thr_B, amelyet a következőképpen számolhatunk:

$$thr_B = \sigma_F + \beta \frac{1}{2N_1} \sum(F), \tag{6}$$

ahol σ_F az adott ablakhosszban lévő értékek átlagos eltérése, β az ablak hossza. Az első küszöbérték azt biztosítja, hogy az adott érték nagyobb, mint a környező területen számított értékek, amelyet egy rövid idejű ablakra számolandó. A második küszöbérték, amelyet egy hosszabb ablakra számolunk, azt

biztosítja, hogy a változás figyelembe veszi az általános tendenciákat az ablakon kívüli adatok változásának figyelembevételével. Az ablakok méretét 3 és 4 másodpercre állítottuk. Ezen küszöbölési technika használata biztosította, hogy a téves elfogadások száma csökkenjen, és csak a valóban magas $KL2$ értéket fogadja el, amelyek a FF-ok határait jelentették.

3. A rendszer kiértékelése

A jelen munka során ugyanazon kiértékelési eljárásokat használtunk, mint ahogyan azt más szegmentálási feladatokban szokás. Az egyik szokásos kiértékelési eljárás a Brandt által kidolgozott GLR módszer [4]. Ez a módszer három gyakori mutatót javasol, amelyek az automatikus szegmentálási teljesítményt mutatják. Az első a beszúrás (Insertion Ins), amely azt jelenti, hogy az automatikus szegmentációban extra határok (események) vannak a referencia annotáláshoz képest. A második mérőszám a törlés (Omission Oms), amely azt jelenti, hogy az automatikus címkesorból hiányoznak események a referencia címkesorhoz képest. A harmadik mérőszám a helyes detektálások száma (Accuracy Acc), amelyet úgy számolunk, hogy a helyesen felismert határok számából – ha az időbeli eltérés az automatikus címkesor és a referencia címkesor között egy előre definiált tolerancia időkeretet nem lép túl – kivonjuk a törlések és beszúrások számát, majd elosztjuk az összes szegmenshatár számával (All):

$$Acc = \frac{Corr - (Ins + Oms)}{All} \tag{7}$$

A helyes detektálások száma fogja legjobban jellemezni a rendszer működését. A rendszer kiértékelése azonban függ a tolerancia időkeret hosszától. A jelen kutatás során több tolerancia értéket vizsgáltunk 25 ms és 100 ms között.

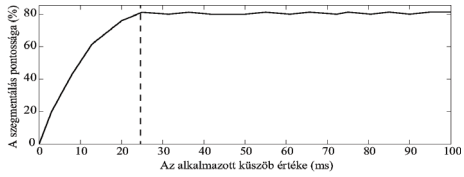
4. Eredmények

Elsőként az IF határok automatikus detektálásának eredményeit ismertetjük. Az eredmények azt mutatják, hogy az általunk kialakított rendszer, amelyben az akusztikai jellemző a beszédjel energiája, a spektrális súlypont és az F0 volt, az IF-ok 83,1%-át jelölte helyesen a spontán beszédben. A legtöbb IF szegmentálási hibát a kitöltött szünetek okozták. A második lépésben az IF-okon belül automatikusan jelöljük a FF-okat. A FF-ok határainak detektálására különböző akusztikai jellemzőket és azok kombinációit teszteltük. Összességében a három akusztikai jellemzővel öt kombinációt teszteltünk. Mindezek mellett változtattuk a $KL2$ érték csúcskeresésében alkalmazott időkeretek hosszát is (100 ms-tól 400 ms-ig). Az eredmények azt mutatták, hogy a helyes detektálások száma jelentős mértékben függ a $KL2$ értékek csúcskeresésében használt időkeret hosszától, illetve az akusztikai jellemzőtől is (1. táblázat). A legjobb eredményt akkor kaptuk, ha csak az alapfrekvenciát használtuk mint akusztikai jellemzőt, illetve a $KL2$ csúcsdetektáláskor használt időkeret hosszát 400 ms-ra állítottuk.

1. táblázat. *FF szegmentálás pontossága a csúcsdetektáláskor használt ablak hosszának függvényében.*

Ablakhossz (ms)	100	200	300	400
F0	68.71	75.83	76.33	**80.18**
Tempó	55.33	56.67	56.91	57.38
F0+energia	68.45	74.75	74.25	79.04
F0+tempó	68.79	73.15	72.33	78.22
F0+energia+tempó	68.86	72.60	71.84	77.05

A FF-ok határainak detektálásakor kiemelten vizsgáltuk a tempó jellemzőt a spontán beszédben, mivel korábbi szakirodalomban bizonyítottan nem volt alkalmazható jellemző a magyar felolvasásokban az FF-ek detektálására olvasott beszédben [21]. Az eredmények szinkronban az előző vizsgálatokkal azt mutatták, hogy ha a tempó jellemzőt önállóan alkalmazzuk a FF-ok határainak detektálására, akkor igen alacsony helyes detektálási arányt kapunk. Ugyanakkor, ha kombináljuk az F0-val, akkor az eredmények javulnak, de még akkor sem érik el azt a helyes detektálási arányt, mint amikor csak az F0 a bemenő jellemző. A helyes detektálási arány abban az esetben is csökken, ha a tempót az alapfrekvenciával és az energiával kombináljuk. Összességében a tempó törlése a jellemzőkészletből a helyes FF-határok detektálását növeli. Ezt az is alátámasztja, hogy a tempó sem az F0-val (R=-0,06), sem az energiával (R=-0,04) nem korrelál. A következő vizsgálatunk a teszteléskor alkalmazott tolerancia időtartamának hatását elemezte a szegmentálás eredményére. Ennek empirikus tesztelésére különböző toleranciatartományokat használtunk (3. ábra). Az eredmények szerint, ha a toleranciatartomány 25 ms, akkor a helyes szegmentálási arány 80,2%. A tolerancia időtartam további emelése már nem hoz eredményjavulást. Mindez azt jelenti, hogy a szegmentáló időben kifejezetten precíznek mondható (hasonló feladatban, igaz, más algoritmussal és más szempontú referenciaképzéssel, de olvasott beszédben 100 ms körül alakult az optimális toleranciatartomány [21]).



3. ábra. A szegmentálás eredménye a tolerancia időtartam függvényében

Az eredményekből arra következtethetünk, hogy az F0 az alapvető akusztikai jellemző a FF-ok detektálásban, az energia pedig a felsőbb szinten, az IF detektálásában fontos jellemző a spontán beszéd esetén.

5. Összegzés

A kutatásban spontán beszéd automatikus prozódiai szegmentálását vizsgáltuk hierarchikus struktúrában, szofisztikált csúcsdetektálási megközelítésben. Első lépésben az intonációs frázist két szünet közötti alapegységnek tekintettük, és k-közép klaszterezéssel detektáltuk. Az IF-okon belül második lépésben FF-ok izolálását kísérletük meg. Az alap prozódiai jellemzőkre (F0, energiaszint és magánhangzó-időtartamok) a korábbi jelértékekhez képesti szimmetrikus Kullback–Leibler-távolságokat számítottuk jeleltolással. A csúcsdetektálást ezzel távolságmetrikával képzett jeleken hajtottuk végre. A három alapjellemző, illetve kombinációik használatával nyert detektálási pontosságot összevetve a legnagyobb pontosságot FF-okra, 80,2%-ot egyedül az F0 használatával kaptuk. Az eredmények alapján az akusztikai markerek tekintetében az energiaszint az IF szintjéhez, az alapfrekvencia pedig a FF szintjéhez sorolható spontán beszédben. Olvasott beszédben a FF-ok detektálásában az F0 domináns, az energiaszint pedig járulékos szerepet játszott [21], spontán beszédben a kísérletek ezt nem mutatták. Az időtartam – legalábbis a magánhangzók hossza – az implementált megközelítésben a FF szintjén nem befolyásolta az eredményeket – ez egybevág az olvasott beszéd esetében tapasztaltakkal. Fontos különbség, hogy a FF-határok a spontán beszédben időben pontosabban detektálhatók, mint felolvasásban (a találati pontosságok felolvasásra ±100 ms, spontán beszédre mindössze ±25 ms időbeli pontosságnál vethetők össze), ennek oka az algoritmikus különbségeken túl a spontán beszéd gyakoribb megszakadása és nagyobb prozódiai dinamikája. Az eredményeket az is befolyásolja, hogy a kiértékelés spontán beszéd esetén percepciós alapú, míg felolvasásra szigorú szintaxis alapon címkézett referenciákkal összevetve történt.

Köszönetnyilvánítás

A kutatás az Országos Tudományos Kutatási Alapprogramok PD112598 számon, "Automatikus fonológiai frázis és prozódiai eseménydetektálás szintaktikai, szemantikai és pragmatikai információk közvetlen kinyerésére a beszédből" címmel támogatott projektje keretében készült.

Hivatkozások

1. Beke A., Szaszák Gy., Váradi V.: Automatic phrase segmentation and clustering in spontaneous speech. In: IEEE 4th International Conference on Cognitive Infocommunications: CogInfoCom 2013, Budapest (2013) 459-462
2. Couvreur, L. Boite, J-M.: Speaker tracking in broadcast audio material in the frame work of the THISL project (1999) 84–89
3. Bolla K.: Szupraszegmentális elemzések. Egyetemi Fonetikai Füzetek 7. ELTE Fonetikai Tanszék, Budapest (1992)
4. Jarifi, S., Pastor, D., Rosec, O.: Brandt's GLR method and refined HMM segmentation for TTS synthesis application (2005) 23–33

5. Chiang, C., Chen, S., Yu, H., Wang Y.: Unsupervised joint prosody labeling and modeling for mandarin speech. Journal of the Acoustical Society of America, Vol. 125, No. 2 (2009) 1164–1183
6. Cruttenden, A.: Intonation. Cambridge University Press (1997)
7. Elekfi L.: Vizsgálatok a hanglejtés megfigyelésének módjaihoz. Nyelvtudományi Értekezések 34. Akadémiai Kiadó, Budapest (1962)
8. Gallwitz F., H. Niemann, E. Nöth, Warnke, W.: Integrated recognition of words and prosodic phrase boundaries. Speech Communication, Vol. 36 (2002) 81–95
9. Giannakopoulos, T.: Study and application of acoustic information for the detection of harmful content, and fusion with visual information. Ph.D. dissertation, Dpt of Informatics and Telecommunications, University of Athens, Greece (2009)
10. Gósy, M.: Virtual sentences in spontaneous speech. In: Speech Research 2003, Budapest, Hungary (2003) 19–43
11. Gósy, M. „BEA - A multifunctional Hungarian spoken language database," PHONETICIAN 105-106, 2012 (2008) 50–61
12. Gussenhoven, C.: The phonology of tone and intonation. Cambridge University Press, Cambridge (2004)
13. Hansson, P.: Prosodic phrasing in spontaneous Swedish. Lund University (2003)
14. Hunyadi, L.: Hungarian sentence prosody and universal grammar. On the phonology – syntax interface. Metalinguistica 13. Peter Lang, Frankfurt/M., Berlin, Bern, Bruxelles, New York, Oxford, Wien (2002)
15. Levelt, W. J. M.: Speaking: From Intention to Articulation. A Bradford Book. The MIT Press, Cambridge, London (1989)
16. Németh T. E.: A szóbeli diskurzusok megnyilatkozáspéldányokra tagolása. Nyelvtudományi Értekezések 142. Akadémiai Kiadó, Budapest (1996)
17. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press (1967) 281–297
18. Siegler, M.A., Jain, U., Raj, B., Stern, R.M.: Automatic segmentation, classification, and clustering of broadcast news audio. In: Proc. of the Speech Recognition Workshop (1997) 97–99
19. Nespor, M., Vogel, I.: Prosodic phonology. Foris Publications, Dordrecht (1986)
20. Olaszy G.: Prozódiai szerkezetek jellemzése a hírfelolvasásban, a mesemondásban, a novella és a reklámok felolvasásában. Beszédkutatás 2005 (2005) 21–50
21. Szaszák Gy., Beke A.: Szintaktikai szerkezet automatikus feltérképezése a beszédjel prozódiai elemzése alapján. In: Tanács A., Vincze V. (szerk.): VIII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2011 (2011) 178–189
22. Yildirim, S., Narayanan, S.: Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, No. 1 (2009) 138–149
23. Varga L.: A hanglejtés. In: Kiefer F. (szerk.): Strukturális magyar nyelvtan 2. Fonológia. Akadémiai Kiadó, Budapest (1994) 468–549
24. Varga L.: The unit of the Hungarian intonation. In: Szathmári, I. (szerk.): Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös nominatae. Sectio Linguistica tomus XXIV. ELTE Eötvös Kiadó, Budapest (1999–2001) 5–13
25. Varga L.: Intonation and stress. Evidence from Hungarian. Palgrave Macmillan, Houndmills, Basingstoke (2002)
26. Vicsi K., Szaszák Gy.: Folyamatos beszéd szó- és frázisszintű automatikus szegmentálása szupraszegmentális jegyek alapján: II. rész: Statisztikai eljárás, finn-magyar nyelvű összehasonlító vizsgálat. In: Alexin Z., Csendes D. (szerk.): III. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2005 (2005) 360–370

27. Wacha I.: Élő nyelvi (spontán) szövegek megnyilatkozásainak (szintaktikai) vizsgálati szempontjaihoz (a gazdagréti kábeltelevízió élő nyelvi felvételei alapján). In: Kontra M. (szerk.): Beszélt nyelvi tanulmányok. Linguistica, Series A, Studia et Dissertationes 1. MTA Nyelvtudományi Intézet, Budapest (1988) 102–158