A korpusz alapú beszédszintézis nyelvi, fonetikai kérdései

OLASZY GÁBOR

BME Távközlési és Médiainformatikai Tanszék
olaszy@tmit.bme.hu

Lektorált

Kulcsszavak: korpusz alapú beszédszintézis, hangszerkezetek, hangátmenetek, optimális vágási pontok, prozódiai invariancia

A beszédprodukció eredménye a beszéd hullámformája. Ez mindenkor egyedi és egyszeri produktum. A beszédszintézis által előállított hullámforma és az emberi produktum közötti az alapvető ellentmondás abban van, hogy a szintézisnél egy tárolt (fix) adatbázisból építjük fel a beszédjelet, tehát megsértjük az egyedi-egyszeri produkcióra vonatkozó tételt. A kérdés az, hogy hogyan lehet ezt az ellentmondást csökkenteni. A korpusz alapú szintézis elvéből fakad, hogy az egyszeri jelzőre vonatkozó időtényezőt próbálja tágítani azzal, hogy hosszabb elemekből építkezik mint a korábbi szintetizálási technológiák, noha itt is egy előre meghatározott, tárolt beszédadatbázisból (korpusz) történik a szintetizált beszéd előállítása. Ennek a törekvésnek a támogatására foglaltuk össze azokat a legalapvetőbb nyelvi, fonetikai ismereteket, amelyekkel segíteni lehet a legjobb jelöltek megtalálását a korpuszban és ezzel a minél jobb hangminőség elérését.

1. Bevezetés

Egy szöveg és annak felolvasott, hangzó formája között szoros összefüggés van. A szöveg a közölt gondolat tartalmát hordozza, valamint tartalmazza a mondat modalitására (kérdés, kijelentés), tagolására utaló írásjeleket is. A szöveg akusztikai szintre való transzformálásakor létrejött hangsorozat képviseli a beszéd úgynevezett szegmentális szerkezetét, (ezt tekinthetjük az írott szöveg tartalmi lényegének, legfőbb hangzó megtestesítőjének). Az előbbi fogalomhoz egy zenei példát adva a kottafejek puszta pontos lejátszása adja a zene szegmentális akusztikai megvalósítását. A valódi értelmezést a művész egyéni megformálása hozza létre az egyéni játékával és a hangszer tulajdonságainak felhasználásával. A beszédben is ez a helyzet, a beszéd „hangszerelését” a hangsúlyozás, a dallammal való variálás, a ritmus megvalósítása és a hangszínezet adja. Ezeket nevezik szupraszegmentális eszközöknek, egy szóval prozódiának.

A beszélőnek a szövegben egyrészről írásjelek jelzik a modalitást, a tagolást, ezek alapján hozza létre a fő dallamszerkezetet, a ritmikát (beszédsebesség, szünettartás, szünethossz). Másrészről a beszélő egy azonnali értelmezést is végrehajt a felolvasásnál. Ez is befolyásolja az előbbi szupraszegmentális elemek variálását. A fenti két nyelvi, fonetikai tényezőt tehát a beszélő automatikusan beleépíti a produkciójába a szegmentális és szupraszegmentális szint egyszerre jelenik meg a jelben. A beszédszintézisben géppel helyettesítjük a beszélő személyt és itt is arra kell törekednünk, hogy az előbbi két nyelvi, fonetikai tényező is bekerüljön az előállított beszédjelbe. A hallgató elvárja, hogy a géppel összeállított beszéd hangzása minél közelebb álljon az emberi hangszínezethez, továbbá, hogy a hangsúlyozás, a ritmikai szerkezet és a beszéddallam kialakítása megfeleljen az adott témának, az ahhoz köthető stílusnak. A beszédszintetizátor produkcióját a hallgató a hangzás alapján értékeli és nem érdekli, hogy a részletek mögött milyen technológia áll. Ezért a legapróbb akusztikai, fonetika részleteket is gondosan tervezni kell, hogy a szintetizált beszédbe minél kevesebb akusztikai hiba kerüljön bele. Ez azt jelenti, hogy nem csupán az akusztikai jellel, mint rezgésformával kell foglalkoznunk a szintézis során, hanem annak nyelvi, fonetikai hátterével is. Ebben a cikkben erre próbálunk rámutatni.

2. A korpusz alapú beszédszintézis fonetika kérdései

Napjainkban a korpusz alapú beszédgenerálás adja a legjobb minőségű szintetizált beszédet.

A technológia nevéből adódik, hogy egy adott beszédkorpusz képezi a szintézis alapját [1] (nagy korpusz, sok órányi beszéddel, annak szöveges és hullámforma-szintű tárolásával és címkézésével). A mérnöki gondolat az így szintetizálandó mondat előállítására az, hogy a korpuszban lépésről lépésre megkeresik az adott szintetizálandó beszédrészleteknek (mondatrész, szófüzér, szó, hangkapcsolat) legjobban megfelelő hullámforma részeket (ezek hossza az esetek többségében lényegesen nagyobb kell hogy legyen, mint egy hangkapcsolat), és azokat kivágva, majd egymás után kapcsolva létrehozzák a mondatot reprezentáló szintetizált hullámformát. Az alapgondolat szerint a cél annak elérése, hogy ne kelljen semmiféle mesterséges jelfeldolgozást végezni a hullámformán, csupán válogatással és összefűzéssel lehessen megoldani az új mondat összeállítását. Az elvárás az, hogy jó válogatás esetén az így előállított beszéd minősége igen közel lesz a korpusz eredeti beszédének a minőségéhez. Ez az új technológia annak a hiányosságnak a ki-

küszöbölésére született meg, hogy az eddigi beszédszintetizáló rendszerek hangminősége ugyan már közel állt az emberéhez, érthető is volt, de közel sem volt olyan hangzású, mint egy ember által felolvasott szöveg, nem lehetett például egyértelműen személyhez kötni, hiányzott belőle a beszélő személy egyéni hangszínezete. A fenti mérnöki gondolatot támogatta még a számítástechnika rohamos fejlődése memóriakapacitásban és feldolgozási sebességben.

A kérdés az, hogy milyen fonetikai, nyelvi szempontokkal kell számolni egy ilyen rendszer tervezésénél és megvalósításánál. A legfontosabb az, hogy figyelembe kell venni a beszédképzés biológiai természetéből adódó tényt, mármint, hogy **az emberi beszéd egyedi és egyszeri produktum**. Az egyedi jelző az egyén saját hangjára vonatkozik, az egyszeri pedig azt fejezi ki, hogy a beszédjel az adott időpillanatra jellemző akusztikai szerkezettel valósul meg. Ez a pillanat a kiejtés pillanata (az akkori biológiai jellemzők határozzák meg a hangot: a gége állapota, a beszélő nyelvi döntései az értelmezést illetően, az artikulációs szervek és környezetük állapota stb.) Még ugyanazon beszélő sem tud egy mondatot ugyanolyan akusztikai szerkezettel kimondani, felolvasni, más lesz a ritmikája, más lesz a hangszínezete, más dallamot valósíthat meg. Az artikuláció tehát pillanatnyi, ugyanakkor folyamatos. Ebből következik, hogy a létrehozott beszédjel is az.

Az említett kiejtési változatosság ellentétben áll azzal a ténnyel, hogy a mai technológiákban a szintézishez a beszédjelet egy rögzített, korpuszból, adott számú tárolt hullámformából, jelválogatással és a jelek összekapcsolásával állítják elő. A rögzített korpusz tartalma mindig ugyanaz, tehát minden esetben ugyanazon korlátozott számú elemhalmaz áll a szintézis rendelkezésére. Az így összeválogatott hullámforma elemek ugyan hangilag tartalmazni fogják a szövegnek megfelelő hangsort, de a beszédjel akusztikai részleteiben a folyamatosság nem jön létre. Olyan eltérések lehetnek az összeillesztési pontoknál (a különböző helyekről való egyedi kivágások következményeként), amelyek megtörik az egyenletes, folyamatos hangzást, ezért zavaróak lesznek a hallgató számára (a beszéd nem lesz természetes hangzású, dallamugrások, hangszínezet-váltások, hangerőváltozások lesznek benne a vágási pontokhoz kötve). Úgy gondolták, hogy a rögzített korpusz nagyságának növelésével ezek a gondok csökkenthetők. Jó példa erre egy Japánban készített korpusz alapú szintetizáló rendszer, amelyik 380 órás hangfelvételű beszédkorpuszból építi fel a beszédjelet [2]. A korpusz növelése sem egyszerű feladat. Számolni kell a beszélő hangszínváltozásával a hosszú felvétel alatt, továbbá a feldolgozási időtényezővel, ami mind az adatbázis elkészítését (egy profi bemondó például 3-4 óránál többet nem tud egy nap teljesíteni, már ez idő alatt is a hangszínezete lényegesen megváltozhat), annak további feldolgozását, előkészítését, mind pedig a későbbi szintézisnél alkalmazott válogatási algoritmust érinti.

3. A vágási pontok teóriája

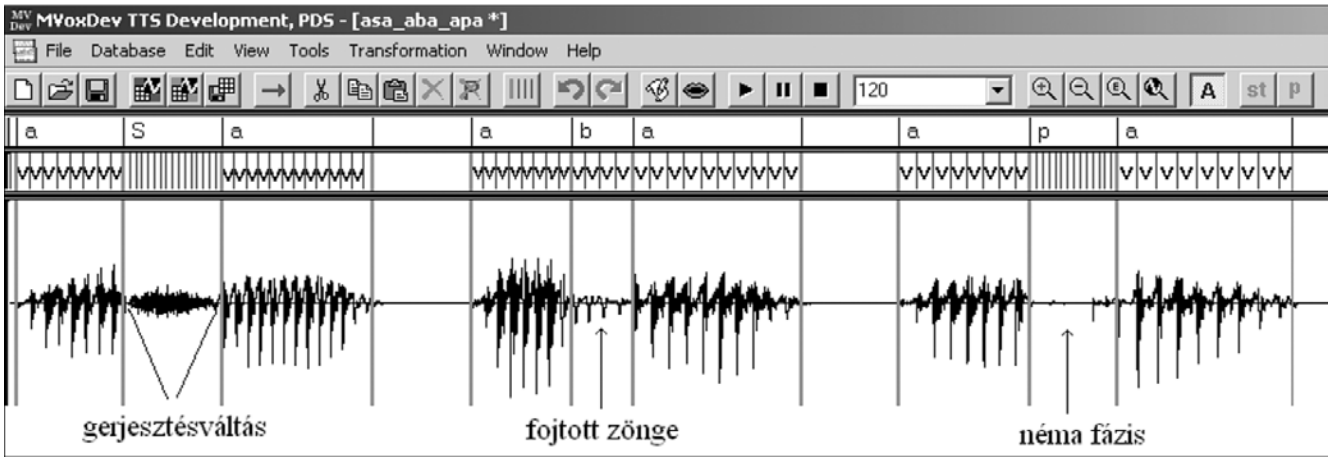
A mai emberi hangú beszédszintetizátorok többsége előre eltárolt, rövid hullámforma-részletekből építkezik, azaz két félhangot (diádot) kapcsolnak össze a hangsor felépítésénél, esetleg kicsivel hosszabb elemet, a két félhang között egy egész hangot tartalmazót, azaz triádokat. A fő kritika a jelenlegi, ilyen elemösszefűzéses technológiákkal szemben az hogy az építőelemek rövidek így az összeillesztési pontok száma az egységnyi (például 1 s-ra számított) beszédjelben sok, ezért sok torzítás kerül a jelbe (minden vágásnál a két összevágott elem határán spektrális eltérések lehetnek, ezek torzítást okoznak). A vágási pontok csökkentésére való törekvés eredménye a korpusz alapú szintézis elve. Ez a teória azonban csak felszíni jelenségre, a vágási pontok számának csökkentésére alapozza megállapítását.

A torzulások az összevágott beszédjelben nem a vágások nagy száma miatt jönnek létre, hanem azért, mert az összevágott jelrészleteket különböző időpontokban ejtett beszédjelből származtattuk, tehát megsértettük azt az előbbi tételt, hogy a beszédjel egyedi és egyszeri. Ha például egy mondat hullámformáját minden hang kezdeténél való szétvágással szétszereljük (sok vágási pontot generálunk), majd újra összeillesztjük, a mondat hangzásában semmilyen minőségromlás nem lesz. Tehát a sok összeillesztési pont önmagában nem kell, hogy generálja a torzítást.

A korpuszos szintézisnél létrejövő minőségjavulás azért jön mégis létre, mert nem rövid jelrészleteket illesztünk egymás után (diádos, triádos hangkapcsolatot), hanem hosszabb beszédszakaszokat (szó, szókapcsolat, szófűzér stb.), tehát ezzel kevésbé sértjük meg a beszédjel egyszeri megvalósulására vonatkozó állapotot. Ha például szavakból, szókapcsolatokból illesztünk össze egy mondatot, akkor az egyszeri, ugyanazon biológiai paraméterekkel megvalósuló beszédprodukciót átfogó időtartam növekszik, tehát hosszabb ideig fogja automatikusan hordozni a beszélőre jellemző hangszínezetet. Ha ehhez még hozzá vesszük, hogy a hallgató hosszabb nyelvi egységekben is kapja ezt a minőséget (szó, szófüzér hosszúságú elemekben), akkor be kell látnunk, hogy a percepció számára kellemesebb, jobb minőségű beszéd jöhet létre, mint amilyen a diádos, triádos rendszereké. Tehát nem az öszszeillesztési pontok számának csökkenése adja a kedvezőbb eredményt, az csak következménye a hangsorépítésnek.

4. Az optimális fizikai vágási pontok meghatározása

A cél, hogy olyan ponton vágjuk el a hullámformát (vegyük ki eredeti korpuszos környezetéből), amely a legkevesebb torzítással jár a későbbi összeillesztésnél. A döntést két tényező befolyásolja: milyen a hangszintű szerkezet az adott ponton és, hogy milyen a prozódiai szerkezet. Itt most a hangszintű szerkezetről fo-



1. ábra Az optimális vágási pontok bemutatása az asa, aba, apa *hármas hangkapcsolatokban. A függőleges vonalak hanghatárok, a zöngés hangok periódushatárait a* v *jelzésű vonalmarkerek jelzik a hullámforma felett.*

gunk elsősorban beszélni. A fizikai vágási pont kialakításához ismerni kell a beszédhangok artikulációs és spektrális belső szerkezetét, valamint tisztában kell lenni a hangkapcsolódások megvalósulásakor létrejövő hangszerkezeti és spektrális módosulások fajtáival. A vágást akkor végezhetjük sikeresen, ha tudjuk, hogy a beszédhangoknak milyen az egymásra hatása, a belső akusztikai szerkezete, hol milyen változás zajlik le a hang frekvencia-, illetve intenzitásszerkezetében a folyamatos artikuláció következtében, melyek azok a hangrészek, amelyek esetleg egymással megegyeznek, illetve nagyon hasonlóak egymáshoz.

A lényeg az, hogy a korpuszból kivágott elemek összeillesztésénél az elemek határán lévő beszédhangokra hangkapcsolódási illesztést kell végrehajtanunk. Úgy kell kiválasztani a kivágandó elemet, hogy ne sértsük meg a spektrális folyamatosság elvét. A következőkben megadjuk a hangsor azon fizikai helyeit, ahol a legoptimálisabbak a vágási pontok: a hangsor minden olyan pontja, ahol gerjesztés váltás megy végbe (tiszta zöngés szakaszt tiszta zöngétlen követ és fordítva, itt ugyanis a jelben intenzitás minimum keletkezik), továbbá a hangok belsejében lévő néma fázisok, zönge szakaszok (ez a zár- és zár-rés hangok sajátja). Az optimális vágási pont kijelöléséhez 5-10 ms pontosságú helymeghatározásra, általában zöngeszinkron jelölésre van szükség *(1. ábra)*. A hangsor összeállításánál ezek után a kivágott beszédrészletek egymáshoz való illesztését hangszerkezeti és artikulációs fonetikai szabályok alkalmazásával tehetjük optimálissá, torzításmentessé. Az illesztés akkor lesz sikeres, ha a beszédjelen nincs hallható akusztikai torzulás a beavatkozás után [3].

Mindezek alapján összeállítottuk a vágási pontokat meghatározó fonetikai szabályokat *(1. táblázat)*. A vágás sikeres elvégzésének alapfeltétele, hogy a hanghatárokat előzőleg pontosan jelöljük be a hullámformán. Az alábbi szabályok minden esetben a hanghatárra, mint vágási pontra vonatkoznak. A vágási pont zöngés hangok esetében vagy a hanghatár, vagy a hang belsejében vonalmarkerrel jelölt periódus határ.

A táblázatból látható, hogy a vágási pont keresését hangszinten kell végrehajtani, ebből adódik, hogy a lexikai formát át kell alakítani fonemikus, hangreprezentációs formává, hogy a szabályok alkalmazhatók legyenek. Ez is fonetikai jellegű tudást igényel.

1. táblázat Példa szabályokra a vágási pontok kijelöléséhez a korpuszban. A hangcsoportok a csatlakozó második hang jelölésére: C= mássalhangzók; V= magánhangzók; C1= p,t,k,ty,h,f, s,sz,c,cs; C2=b,d,g,gy,zs,z,dz,dzs; C3= v, j, l, r ; C4= m, n, ny; kiv:= kivéve

megelőző hang a betűjele szerint	következő, kapcsolódó hang a betűjele szerint	vágási pont kijelölésének szabálya	szöveges példa (a csatlakozó hangokat kiemeléssel jelöltük)
a) b, d, g, gy b) b, d, g, gy	a) V, C3, C4 b) önmagával csatlakozik	a) a hanghatáránál kell vágni b) a hosszú hang 70%-ánál kell elvágni, a zárfelpattanás nem lesz benne)	*vad vihar* *nagy meleg* *vad dörrenés*
n	a) V, C kiv. C4 b) önmagával	a) a hang határánál kell vágni b) a hang 70%-ánál kell elvágni	*télen derült* *télen nagyon*
f, sz, s, c, cs, h	a) V, C3, C4, C1 kiv. b), b) önmagával	a) a hang határánál kell vágni b) a hang 70%-ánál kell elvágni	*havas lesz* *havas síkos.*
l	a) V b) C kiv. c) c) önmagával	a) nem célszerű elvágni b) a hang határnál kell elvágni c) a hang 70%-ánál lehet elvágni	*szél óránként* *tél marad* *fel lesz*

5. Az artikuláció akusztikus vetülete

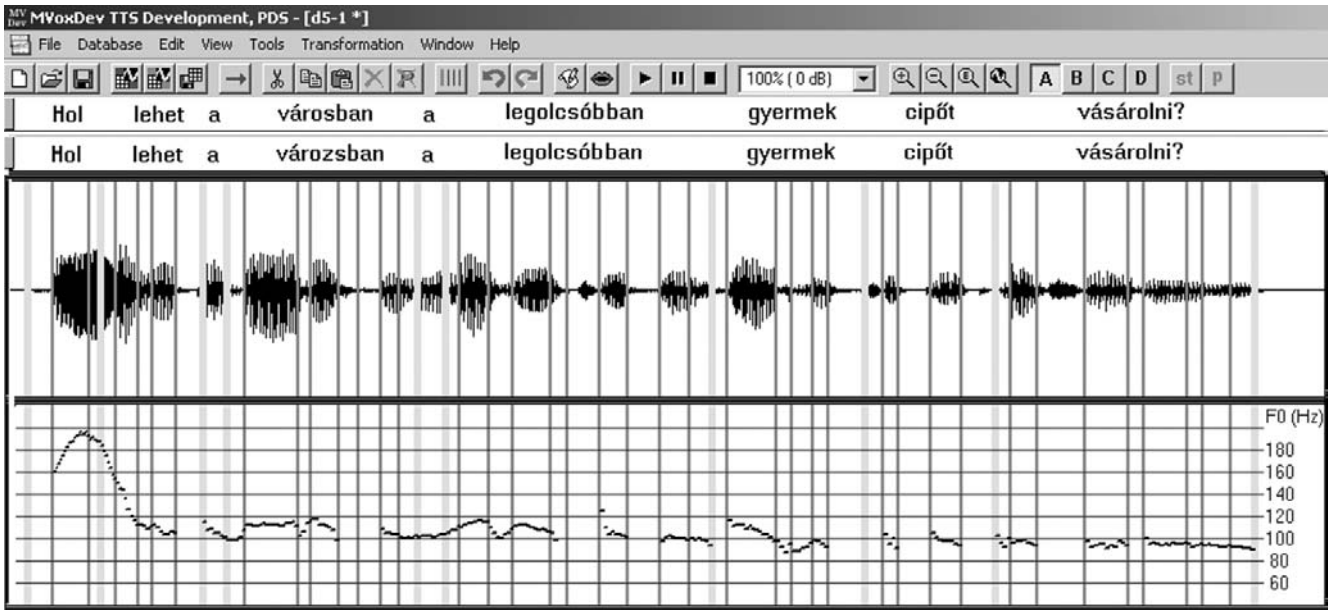
Amennyiben a táblázat szabályai szerinti hangkombinációt a korpusz nem tartalmaz, akkor a torzításmentes vágási pont megtalálásához másodlagos jelölteket is lehet állítani. A megoldás elméleti hátterét a hangok képzésekor kialakuló artikulációs konfiguráció ismerete adja [3]. A képzési hely önmagában reprezentál egy elméleti akusztikai tartalmat. Minden artikulációs pozíciónak megvan a saját statikus spektrális megfelelője az akusztikai térben, amit a formánsok, illetve zörejgócok frekvenciáival fejeznek ki. A folyamatos beszédben a képzési konfigurációk (mozgássorozatok) követik egymást, a mozgássorozatok egymásra hatását pedig akusztikai szinten a hangok úgynevezett átmeneti fázisai tartalmazzák (a formánsok mozognak). A másodlagos jelöltként felhasználható vágási szabályokat CV és VC elemekre vonatkoztatjuk. A cél az egymáshoz hasonló, így egymással helyettesíthető hangkapcsolati elemek meghatározása. A helyettesítő szabályok kidolgozásánál a mássalhangzók képzési helyeiből és azok akusztikai tartalmából, mint statikus tényezőkből indulunk ki, majd a CV, VC összekapcsolódásból adódó dinamikus változásokat tanulmányozzuk, vagyis azt, hogy a mássalhangzók hogyan hatnak a magánhangzók spektrális szerkezetére.

Akusztikai szempontból kitüntetett szerep jut itt az azonos, illetve közel azonos képzési helyű mássalhangzóknak, mivel azok közel azonos statikus akusztikai hatással rendelkeznek, ezért azonos hatást fejtenek ki a magánhangzóra, annak a mássalhangzóhoz csatolódó részére, az átmeneti fázisára (például az *ások, ácsok* szavakban az (o)-ban szinte ugyanaz a spektrális szerkezet alakul ki a (s), illetve a (cs) hatására). Másrészről ezen mássalhangzók hangzó része is egymáshoz hasonló spektrális szerkezetű, ha a gerjesztést nem változtatjuk meg (például az ások, ácsok szavakban az (s), illetve a (cs) hang spektrális komponensei igen hasonlóak). A hangkapcsolódásra jellemző spektrális szerkezet tehát előre ismerhető (a vágási pont optimalitási szintje jósolható). A magyar mássalhangzók képzési hely szerinti csoportosítását a *2. ábra* mutatja. Az azonos képzési helyű hangok egy-egy sort képviselnek, ezek a sorok képezik a másodlagos vágási pontok kijelölésének az alapját. Az előbbi okfejtés szerint tehát az ábra minden vízszintes sora egy-egy artikulációs pozíciót reprezentál. Ennek a pozíciónak az akusztikai vetülete elméletileg hangtól függetlenül átkerül a mássalhangzót követő, illetve megelőző magánhangzóba. Ez azt jelenti, hogy az azonos sorban szereplő mássalhangzók hatása a magánhangzók formánsszerkezetére jó közelítésben azonos. Mivel a nazális mássalhangzók többnyire nazalizálják a magánhangzót is, ezeket nem célszerű a többi ugyanazonsori mássalhangzóval egyforma kategóriába sorolni.

A 2. ábra első és második sorában tehát gyakorlatilag csak a (b, p), illetve a (v, f) hangpár használható vágási szabályhoz. A harmadik sorban hét hang, a negyedikben hat található. Tehát ez a két artikulációs pozíció viszonylag széles körben használható a helyettesítésre. Az ötödik sorban szereplő palatális mássalhangzók pedig képzési helyük kitüntetett volta miatt használhatók jó hatásfokkal (a palatális mássalhangzók erős és jellegzetes hatást gyakorolnak a magánhangzók formánsszerkezetére [4,5]). A hatodik sorban látható (g, k) mássalhangzókat pedig alkalmazkodó képességük miatt lehet hatásosan felhasználni. Ezek a mássalhangzók ugyanis a legképlékenyebbek, a képzési helyük változik a hozzájuk kapcsolódó magánhangzó függvényében. Lássunk a fentiek alkalmazására egy példát. Ha a szöveg szerint *szá* kapcsolat hanghatárán kell illeszteni (például az *sz* az előző szó utolsó eleme az *á* pedig a hozzá csatolandó szó első hangja), de nincsen a korpuszban csak *szo*, viszont van *cá*, vagy *tá*, akkor az utóbbi kettő felhasználásával kiválasztható a helyettesítő elem. A 2. ábra szerint az (sz) hangot helyettesítheti a (c), illetve a (t) hang is, így a helyettesítő kapcsolódó elem lehet a *tá* vagy *cá* kapcsolatból kiválasztott (á) hang is. Az így összeillesztett (sz)+(á) hullámforma nem tartalmaz spektrális torzítást, mivel a fenti mássalhangzók a magánhangzóban ugyanazt a spektrális átmeneti fázist hozzák létre.

2. ábra A magyar mássalhangzók képzési hely szerinti csoportosítása

	zárhangok								zár-réshangok				réshangok							nazálisok					
	b	p	d	t	gy	ty	g	k	c	dz	cs	dzs	v	f	z	sz	zs	s	h	m	n	ny	j	l	r
két ajak	☒	☒																		☒					
ajak-fog													☒	☒											
fog-fogmeder			☒	☒					☒	☒					☒	☒					☒				
fogmeder											☒	☒					☒	☒						☒	☒
elülső szájpadlás					☒	☒																☒	☒		
hátsó szájpadlás							☒	☒																	
gége																			☒						



3. ábra Egy kérdő mondat Fo menetének vizsgálata a szöveg függvényében. A hanghatárokat vékony, a szóhatárokat vastag függőleges vonalak jelzik.

6. A prozódiai tartalom és a szöveg összefüggései

A vágási pont optimális meghatározását – mint korábban említettük – két tényező befolyásolja: milyen a hangszintű szerkezet az adott ponton és hogy milyen az alapfrekvencia (Fo=hangmagasság) és az intenzitás (I) a kapcsolati ponton (a prozódia két fő eleme). A kérdés azért fontos, mert a hangsorban nem csak a hangfolyam képez egy folyamatos egységet, de a prozódiai szerkezet elemei is. A prozódiai szerkezet lefolyása a mondat elejétől a végéig folyamatos és jellemző a mondatra. A prozódiával tehát külön is kell foglalkoznunk, hogy az összekapcsolt hullámforma részletek ilyen szempontból is illeszkedjenek egymáshoz. Hogyan határozhatunk meg olyan vágási pontokat, amelyek a hangszíntű optimalitáson túl a prozódiai folytonosságot is biztosítják? Hogyan jósolható a prozódia a szövegből?

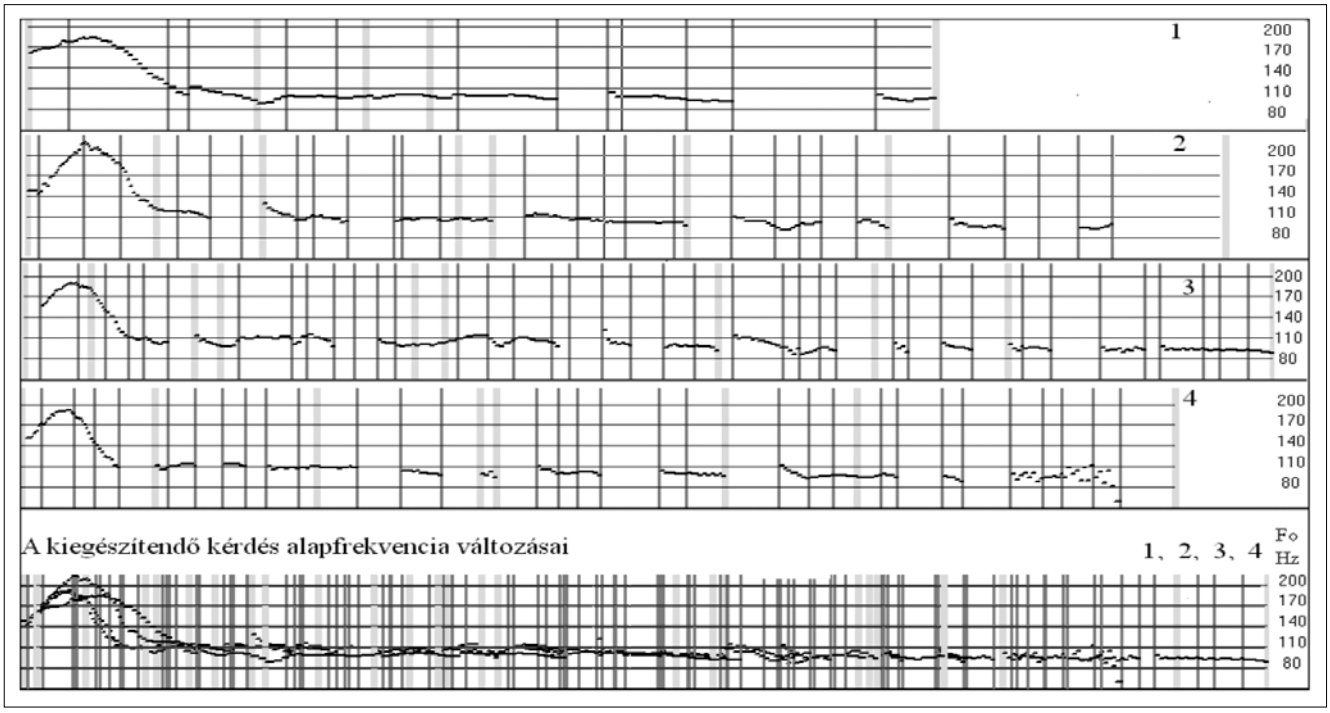
A problémakört két irányból közelíthetjük meg. Az egyik, amikor a szöveg oldaláról végzünk nyelvészeti elemzést, és a szövegben megjelöljük a várható fő dallamformákat, kijelöljük a hangsúlyokat, meghatározzuk a szünetek helyét, és azt mondjuk, hogy a felolvasó személy nagy valószínűséggel ezek szerint fogja felolvasni a szöveget (vö. [6]). Az elemzés eredményét rávetítve a felolvasott beszédjel Fo görbéjére, intenzitás függvényére megállapíthatjuk, hogy vannak-e eltérések a nyelvészeti elemzés és a produkció között. Ha nincsenek jelentős eltérések, akkor a nyelvi elemzés segítheti a helyes prozódiai szerkezetek azonosítását a beszédkorpuszban a szöveg-hullámforma megfeleltetésnél.

A másik megközelítés fonetikai jellegű. Itt a beszédprodukció oldaláról indulunk ki és a fizikai paraméterek változását vetítjük rá a kiindulási szövegre, majd a megjelölt változási pontok alapján vonunk le következtetéseket a szöveg és a prozódia kapcsolatáról. A következőkben erre mutatunk példát. A vizsgálatunk középpontjában a beszédjel prozódiai szerkezete áll, és az, hogy ennek a szerkezetnek a változási pontjait rávetítsük a szövegtestre. A kérdés tömören az, hogy mennyire invariánsak a magyar mondat dallamformái és intenzitás függvényei. Ha sikerül a változási pontokat mind a szövegtestben, mind a prozódiai elemekben egymással szövegfüggetlenül összerendelni, akkor az adott szövegrész megváltoztatásával és az arra vonatkozó prozódiai szerkezet megtartásával elérhetjük, hogy a beszéd tartalmi része megváltoztatható (más szövegrésszel kicserélhető a két változási pont között) anélkül, hogy a prozódiai hangzásban minőségi csökkenés következne be. Ez azt eredményezi, hogy a korpuszban úgy kereshetünk szöveget, hogy meg tudjuk jósolni a dallammenetét és hangsúlyozását is.

A magyar beszéd prozódiai (szupraszegmentális) szerkezetével kapcsolatos eddigi elemzések, valamint az egyes részterületekre korlátozott modellezési formák jó kiindulási alapot szolgáltatnak a fenti vizsgálatok végzéséhez [7].

7. A prozódiai vizsgálatok anyaga és módszere

A vizsgálatokhoz olyan beszédadatbázisokat használtunk fel, amelyekben a bemeneti szövegtest mellett szerepeltettük a fonemikus átírási formát is, valamint ezzel párhuzamosan tároltuk az elhangzott beszédjelet, annak hang- és szó szintű címkéit, valamint az ezekhez tartozó időtartam, alapfrekvencia és intenzitás adatokat *(3. ábra)*. A jelen vizsgálatnál fontos szempont, hogy ugyanazon beszélő hangját vizsgáljuk és az összehasonlításokat is egyetlen hangra kell vonatkoztatni, hiszen a korpusz alapú szintézisnél is egy adott beszélő hangjából akarunk szintetizált mondatokat előállítani.



4. ábra Négy kiegészítendő kérdés egyenkénti Fo menete, és ezek egymásra helyezett képei (alul). A függőleges vékony vonalak a hanghatárokat jelölik, a vastagabb szürke vonalak pedig a szóhatárokat.

A célkitűzés megvalósítására egyszerű kijelentő és kérdő mondatok alapfrekvencia- és intenzitásszerkezetét elemeztük A mintamondatokat egy olyan beszédadatbázisból válogattuk amelyben egy férfi bemondó hangját rögzítettük [8]. A kísérlethez 16 kijelentő és 8 kérdő mintamondatot válogattunk ki az adatbázisból. A mintamondatok mindegyikén jellemeztük az alapfrekvencia változást annak töréspontjaival, valamint az intenzitások alakulását. A mondatok szövegtartalmát figyelmen kívül hagytuk a vizsgálat során, mivel mondat szinten voltunk kíváncsiak a szövegtől független Fo és I szerkezetek alakulására.

8. A prozódiai vizsgálatok eredményei

8.1. A kérdések vizsgálata

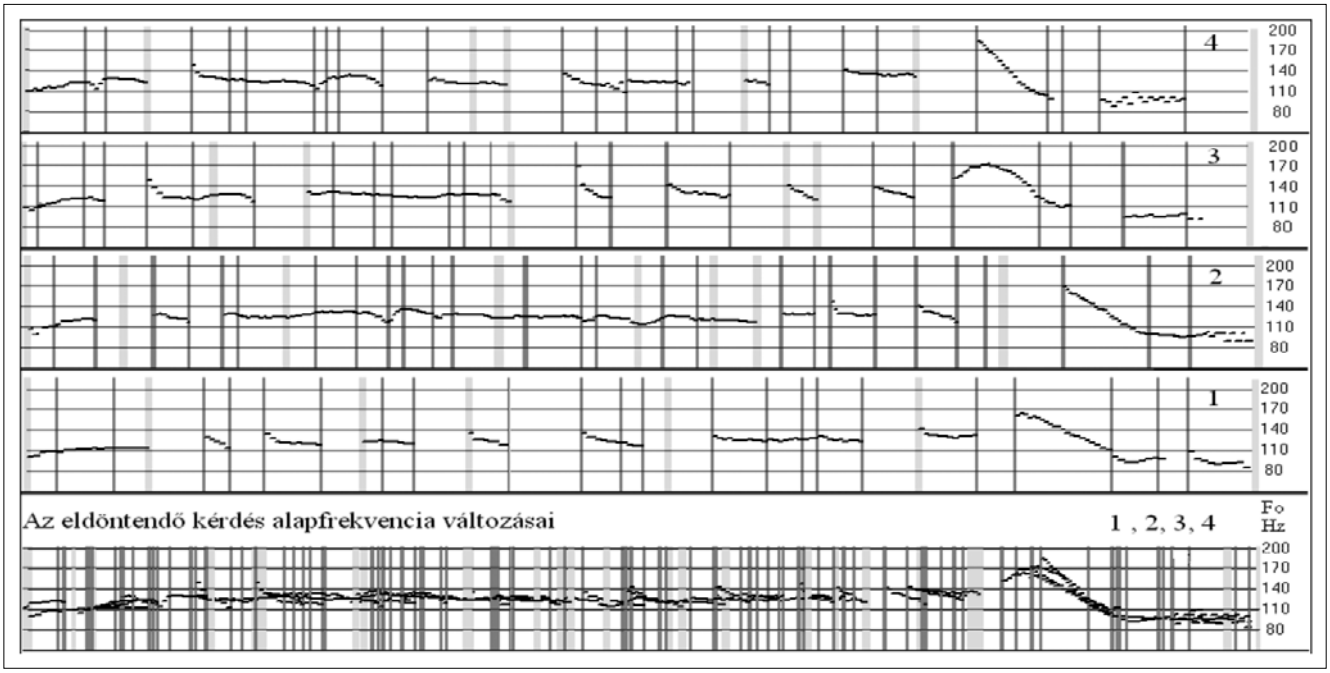
A mért kiegészítendő kérdésekben az alapfrekvencia mozgása a kérdésre jellemző jól ismert képeket mutatja, a vizsgálat tárgya itt a prozódiai szerkezet stabilitása. Az összesített vizsgálati eredmények azt mutatják, hogy a kérdésekre produkált alapfrekvencia szerkezetek egységes képet mutatnak. A kérdés magján a kérdőszón van az intonációs csúcs (meredek felfutás és meredek csökkenés), utána pedig enyhén esik az Fo.

Részletezve az előbbi leírást a vizsgált bemondóra azt mondhatjuk, hogy az Fo csúcs a kérdőszó első szótagjának magánhangzóján 200 Hz körüli értéken van. Ezután erőteljes meredek esés következik a második szótag magánhangzójának a végéig, itt 110 Hz körüli az Fo. Ettől a ponttól kezdve az alapfrekvencia folyamatosan csökken 90 Hz körüli értékre a mondat végére. A mondatok tartalmi része nem befolyásolja a dallammenetet, az ugyanazon bemondó kérdéseiből átlagolt görbe mentén kicsi az Fo szórása *(4. ábra)*. Ez azt jelenti, hogy a kiegészítendő kérdés dallammenete a vizsgált beszélőre három jellemző ponttal leírható. Az intenzitás alakulására ugyanez mondható el. Ezeket a pontokat a szöveg szintjére vetítve megkapjuk az általános sémát a kiegészítendő kérdések és a szöveg kapcsolatáról. A fenti három pont helyzetének szövegszintű azonosításához csak a szótagok helyzeti meghatározására van szükség (szó szintű információt nem használunk). A jellemző pontok a következők: az első szótag magánhangzója, a második szótagé, továbbá a mondat utolsó szótagja.

Az eldöntendő kérdésekben az Fo menet szintén stabil képet mutat, független a mondat tartalmától *(5. ábra)*. A csúcs a kérdés magján van, az utolsó előtti szótagon. Részletezve: a mondat indítására a 110 Hz körüli érték a jellemző, fokozatos emelkedés következik 140 Hz-re az utolsó előtti szótag elejéig, majd hirtelen ugrás következik 170 Hz körüli értékre amit az Fo a szótagmag elején ér el, majd a szótag végéig hirtelen csökkenés következik be a 90 Hz körüli értékre. Ez az érték marad a mondat végéig. Mint látható itt négy jellegzetes pontot lehet kijelölni a szövegtől függetlenül: a mondat kezdete, az utolsó előtti szótag eleje, ezen belül a magánhangzó eleje, majd a magánhangzó vége.

A kérdések tekintetében tehát azt az összegzett következtetést vonhatjuk le, hogy az alapfrekvencia görbe egységes, invariáns képet mutat és a szövegtől függetlenül jellemző a kérdésre.

A fentiekből látható, hogy a kérdésekben a szövegre vetített prozódiai információt a szöveg szótag szerkezetéhez lehet kötni, sem a szavak sem a szöveg tar-



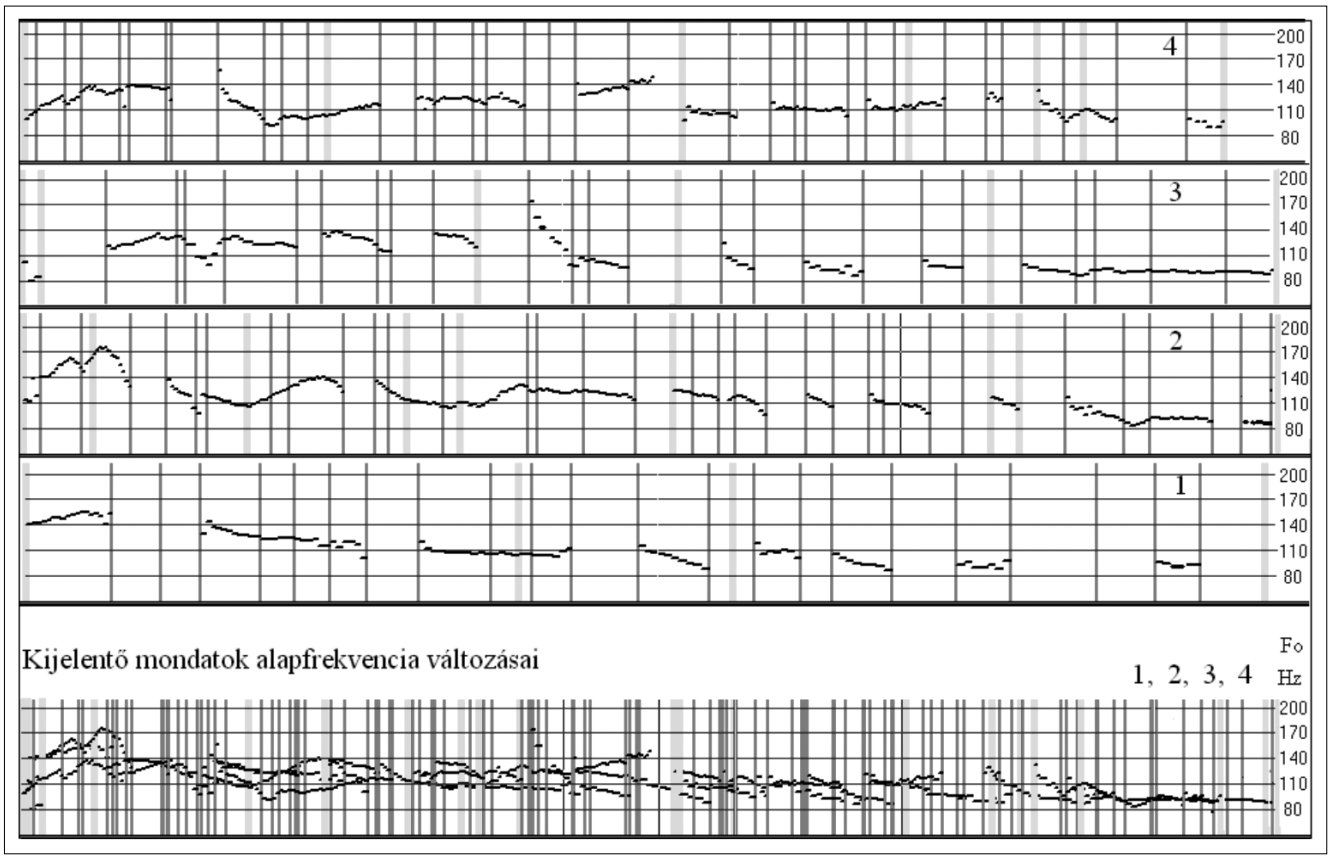
5. ábra Négy eldöntendő kérdés egyenkénti Fo menete, és ezek egymásra helyezett képei (alul)

talma nem befolyásolja a prozódiai szerkezet kialakítását. Mindezekből azt a következtetést vonhatjuk le, hogy az egyszerű kérdések előállításához nincs szükség a szöveg semmilyen elemzésére, hiszen az Fo és I töréspontok egyértelműen, szótaghoz kötötten meghatározhatók. Ez megkönnyíti a válogatást a korpuszban.

8.2. A kijelentő mondatok vizsgálata

A vizsgált mondatokra egyenként végeztük el az Fo menet elemzését és a jellemző pontok hozzárendelését a szöveghez. Az elemzések eredményei a következők. A kijelentő mondat Fo menetében változást okoz a mondathangsúly helye, a szó hangsúlyos volta, a hangsúlyos szavak helye a mondatban, valamint a határjel-

6. ábra Négy kijelentő mondat egyenkénti Fo menete, és ezek egymásra helyezett képei (alul)



zések (vessző, gondolatjel, pontosvessző, kettőspont). Úgy találtuk, hogy az Fo általában a mondat első hangsúlyos szótagján a legmagasabb értékű, de ez attól is függ, hogy a beszélő hogyan értelmezi a mondatot a hangsúlyok vonatkozásában. Amennyiben van mondathangsúly, akkor az lehet a legmagasabb értékű. A hangsúlyos szavak első szótagján Fo emelkedés található, a visszacsökkenés az esetek nagy többségében a második szótagban zajlik le. Az Fo csúcs kiemelkedésének a mértéke a hangsúlyos szótagban általában függ a szó mondatbeli helyzetétől. Minél távolabb vagyunk a mondat elejétől, annál jobban csökken az Fo kiemelkedése a környezetéből.

Az utolsó szavakban (ha ezek hangsúlyosak is) ez a kiemelkedés szinte csak pár Hz. A hangsúlyok közötti részeken az Fo enyhe esést mutat, azonban ezt az eső tendenciát megváltoztathatja a határjelzés, illetve a mondathangsúly. Ilyenkor nem eséssel kell számolni, hanem szinten tartással, esetleg enyhe Fo emelkedéssel. A kijelentő mondatok alapfrekvencia szerkezete tehát sokkal bonyolultabb képet mutat, mint a kérdéseké. A *6. ábrán* a vizsgált anyagból négy mondat Fo szerkezetét mutatjuk be.

Az összesített képből látható, hogy az Fo szórása sokkal nagyobb, mint a kérdéseknél. A mondat belsejében nem lehet jellemző Fo karakterisztikát találni. Ez a mondat belseji hangsúlyok más-más elhelyezkedéséből fakad. A mondat elejére ki lehet mondani, hogy az magasabb Fo-al rendelkezik, mint a mondat vége. Az egyedüli egységes pont, ami minden ilyen kijelentő mondatra jellemző a mondatvég alapfrekvenciájának értéke, az Fo 85-90 Hz-re csökken (a vizsgált bemondó ejtésében). Az intenzitásszerkezeti kép egységesebb, mint az alapfrekvencia. A mondat kezdetén kialakuló intenzitás jellemző a mondat nagy részére, a befejező szakaszban az intenzitás csökken.

Látható tehát, hogy a kijelentő mondatokban a prozódia és a szöveg kapcsolatának kijelölése bonyolult szövegelemzést is igényelhet, hogy a megfelelő prozódiai részeket függetlenítsük a szöveg tartalmától. Ez nem kedvező eredmény a korpusz alapú szintézis szempontjából, hiszen a legtöbb esetben kijelentő mondatokat, közléseket kell előállítani. A probléma kezeléséhez az szükséges, hogy csökkentsük a mondatok variáltságát a korpuszban. Ezért a korpusz alapú rendszereket ma még csak meghatározott témakörökre (például időjárás jelentés, jegyrendelés) fejlesztenek. Itt elérhető, hogy a korpuszban viszonylag állandó szerkezetű mondatokat tárolunk és ezekből építjük fel az új mondatot. Ennek köszönhető, hogy a nyelvi elemzésnél egyszerűbb módszerekkel is modellezni tudják a kapcsolatot a kijelentő mondat prozódiai szerkezete és a szintetizálandó mondat szövegteste között (lásd Fék és tsai. cikkében, ugyanebben a számban).

9. Összefoglalás

A korpusz alapú beszédszintézis-technológia nyelvészeti-fonetikai kérdéseit tárgyaltuk. Rámutattunk arra, hogy a beszédjel akusztikai megjelenéséhez szorosan hozzátartozik a fonetikai, nyelvi háttér is. A jó akusztikai végeredmény (emberi hangú szintetizált mondat kellemes hangszínezettel és hanglejtéssel) eléréséhez ezeket az ismereteket is fel kell használni a korpusz alapú szintetizáló rendszerek tervezésénél.

Köszönetnyilvánítás

Ezt a kutatást az NKFP 2. programja (szerződésszám: 2/034/2004) támogatta.

Irodalom

[1] Schweitzer A., Braunschweiler N., Klankert T., Möbius B., Sauberlich B., Restricted Unlimited Domain Synthesis. Proc. Eurospeech 2003, Geneve, pp.1321–1324.

[2] Kawai H., Toda T., Ni J., Tsuzaki., Tokuda K., Ximera: a new TTS from ATR based on corpus-based technologies. Proc. of the 5th ISCA Speech Synthesis Workshop, Pittsburgh 2004.

[3] Olaszy Gábor, Az artikuláció akusztikai vetülete – a hangsebészet elmélete és gyakorlata. Kif-LAF 2003. Szerk.: Hunyadi László, Debreceni Egyetem 2003, pp.241–254.

[4] Magdics Klára, A magyar beszédhangok akusztikai szerkezete. NytudÉrt. 49, Akadémiai Kiadó, Budapest, 1965.

[5] Olaszy Gábor, A magyar beszéd leggyakoribb hangsorépítő elemeinek szerkezete és szintézise. NytudÉrt. 121, Bp., 1985.

[6] Tamm Anne, Olaszy Gábor, Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához. In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szerk.: Alexin Zoltán és Csendes Dóra, Szegedi Tudományegy. Informatikai Tanszékcsoport, Szeged 2005, pp.383–393.

[7] Olaszy G., The most important prosody patterns of Hungarian. Acta Linguistica Hungarica, Vol. 49 (3-4), 2002. pp.277–306.

[8] Olaszy Gábor, Abari Kálmán, Adatbázisok és számítógépprogramok a magyar beszéd időszerkezeti vizsgálatához. Alkalmazott Nyelvtudomány 2., 2005, pp.41–62.