Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Távközlési és Médiainformatikai Tanszék

A gépi beszéd-előállítás természetességének növelése rejtett Markov-modell alapú szövegfelolvasó rendszerben

Ph.D. tézisfüzet
BME-VIK Informatikai Tudományok Doktori Iskola

Csapó Tamás Gábor
okl. mérnök-informatikus

Témavezető:
Németh Géza, Ph.D.

Budapest, 2013

1. Bevezetés

Az információs társadalomban az ember-gép kapcsolat kutatásába illeszkedik a beszéd gépi előállításának minél jobb minőségű megvalósítása. A felhasználó és a gép között beszéd segítségével megvalósuló kommunikáció igen fontos, ha a felhasználó keze és látása lekötött (pl. autóvezetés közben), illetve sérülés miatt nem használható (pl. látássérültek), továbbá ha az igénybe vett szolgáltatás telefonvonalon keresztül érhető el (pl. intelligens tudakozó, hírolvasás mobil eszközön). Az expresszív, érzelmeket imitáló gépi beszéd akkor lehet előnyös, ha hosszabb szöveg felolvasásában szeretnénk a monotonitást csökkenteni (pl. hangoskönyvek esetén). Az adott beszélő hangján megszólaló, személyre szabott gépi szövegfelolvasó rendszerek hasznosak lehetnek azon felhasználóknak is, akik sérülés vagy betegség miatt elvesztették hangképzési lehetőségüket.

A beszéd képzésének számos egyszerűsített modelljét hozták létre, melyek nagyrészt a forrás-szűrő szétválasztáson alapulnak [1]. A gégének, vagyis annak a hangképző szervnek, amit forrásnak tekintünk, durva modellje lehet akár egy egyszerű impulzussorozat a zöngés szakaszokban és fehér zaj a zöngétlen részeken. A toldalékcső (szájüreg, orrüreg, stb.), azaz a szűrő modellezésére is sokféle eljárást dolgoztak ki. A gépi szövegfelolvasás egyik legújabb technológiája, a statisztikai parametrikus beszédszintézis is sok esetben a forrás-szűrő modellt használja [2]. A toldalékcső modellezése már elérte azt a szintet, ahol a további minőség javítás csak nagy befektetett energiával érhető el és a kutatás nem ezen a ponton kritikus [3]. A forrásjel modellezésére azonban még nem született kiforrott technika, melynek segítségével a statisztikai parametrikus beszédszintézis hangkarakterisztikája általános körülmények között is elérné az elemkiválasztásos rendszerek[1] [4] nyújtotta természetességet. A forrás modellezése ma is aktív kutatási terület.

A legtöbb beszédtechnológiai módszert idealizált beszéd feldolgozására készítették el. Ideális zöngés beszédet feltételezve a hangszalagok kváziperiodikus módon rezegnek, azaz az egyes zöngeperiódusok között csak kis változások figyelhetők meg. A természetes beszédben azonban a beszélők időnként ettől különböző zöngeképzéssel beszélnek, és a beszédjelben az ideálistól lényegesen eltérő jellegzetességű (pl. kiugró vagy erősen lecsökkent amplitúdójú) zöngeperiódusok is megfigyelhetők. Ugyan már léteznek módszerek ezen jelenségek elemzésére, detektálására és transzformációjára [5], de az ideálistól eltérő beszéd (pl. irreguláris zöngeképzés) szintézisben történő modellezésével és az ehhez kapcsolódó transzformációs eljárásokkal keveset foglalkoztak.

[1] Az elemkiválasztásos beszédszintézis lényege, hogy az élő személy hangjának rögzítésével kialakított beszédkorpuszból minél hosszabb elemeket (szavakat, szókapcsolatokat) egymás után fűzve próbálja meg a szöveghez tartozó beszédet előállítani.

A fenti forrás-szűrő szétválasztáson alapuló modellek azt feltételezik, hogy a forrás és a szűrő tökéletesen szétválasztható az emberi beszédkeltés során. Azonban ez nem mindig teljesül, és nemlineáris csatolás jöhet létre a forrás és a szűrő közötti interakció miatt. Az utóbbi néhány évben kimutatták, hogy a gége és a felette lévő szervek mellett az alsó légúti rendszer (pl. tüdő, légcső, hörgők) is befolyásolja a beszédet [6]. Eszerint az alsó légúti (szubglottális, azaz gége alatti) rendszer hozzájárul a magánhangzók megkülönböztető jegyek szerinti elkülönüléséhez [7], azaz szerepet játszik a beszédhangok egymástól való megkülönböztetésében. Az alsó légúti rezonanciák beszédtechnológiai felhasználási lehetőségeit eddig csak kezdeti kísérletekben vizsgálták.

A téziseimet a fenti témáknak megfelelően három csoportra bontottam. Az I. téziscsoportban bemutatok egy újszerű beszéd gerjesztési modellt, és ismertetek egy ezen alapuló irreguláris-reguláris beszéd transzformációs eljárást. A II. téziscsoportban a modell beszédszintézis vonatkozásait dolgozom ki és bemutatok két új irreguláris zöngeképzési modellt, amelyek statisztikai parametrikus szövegfelolvasóban használhatóak. A III. téziscsoportban a toldalékcső és az alsó légúti rendszer közötti kölcsönhatással foglalkozom: a szubglottális rezonanciák vizsgálatára irányuló kutatásomat ismertetem.

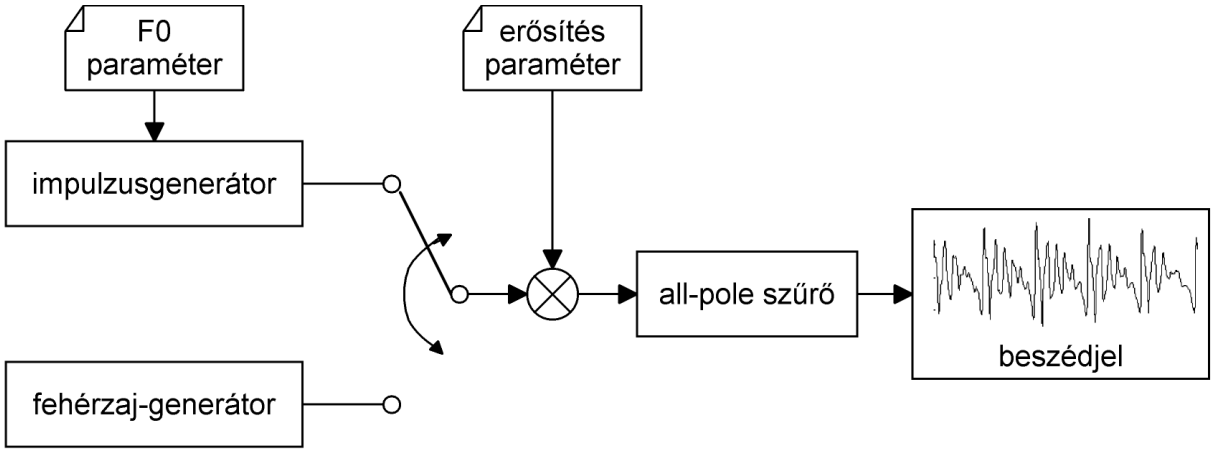
2. Irodalmi áttekintés

A legkorszerűbb beszédszintézis technológiák egyike a rejtett Markov-modell alapú beszédszintézis (*Hidden Markov-model based Text-to-Speech*, HMM-TTS), amely a statisztikai parametrikus szövegfelolvasók családjába sorolható [3]. Ennek egyik kutatási eszköze a nyílt forráskódú HTS rendszer [2]. Az elmúlt években a HMM-TTS nagy népszerűségre tett szert számos előnyös tulajdonsága miatt: flexibilis, alacsony memóriaigényű és nem tartalmaz olyan zavaró akusztikai torzításokat, mint a korábbi elemkiválasztásos rendszerek [3].

A statisztikai parametrikus beszédszintézis során nem közvetlenül a beszédadatbázis hullámformáin végzünk átalakításokat, hanem a beszédet először paraméterekre bontjuk, amelyeket gépi tanuló algoritmus kezel a továbbiakban. A tipikusan néhány óra hosszúságú beszédkorpuszból először a gerjesztési és spektrális paramétereket nyerjük ki. Ezen paraméterek felhasználásával megtörténik a HMM-ek tanítása a beszédkorpusz fonetikus átirata és a beszédhangokra vonatkozó környezetfüggő címkék alapján. A tanítás eredménye a kisméretű HMM adatbázis, amely a szintézisben használható fel. A szintézis során a felolvasandó szöveghez először fonetikus átirat és környezetfüggő címkézés készül, majd a HMM adatbázis alapján a címkézett szöveghez megfelelő paramétereket generál a rendszer. A generált gerjesztési és spektrális paraméterek alapján egy beszéd visszaállító eljárással készül a beszéd hullámforma [8].

2.1. Gerjesztési modellek a statisztikai parametrikus beszédszintézisben

A legtöbb HMM-TTS rendszer a beszéd forrás-szűrő szétválasztásán alapul [1]. A legegyszerűbb, impulzus-zaj módszerre az 1. ábra mutat példát: a zöngés szakaszokat alapfrekvencia-függő ($F0$) impulzussorozattal, a zöngétlen részeket sávkorlátozott fehér zajjal modellezzük, majd az összefűzés és erősítés után all-pole szűréssel kapjuk meg a beszédjelet. Az alap HTS rendszerben lévő egyszerű impulzus-zaj gerjesztés azonban a HMM-TTS rendszer minőségét „zizegőssé", robotossá teszi az elemkiválasztásos rendszerek tiszta, csengő hangjához képest (zizegős beszéden az egyszerű beszédkódolók által eredményezett fémes, gépies, robotos hangot értem; angol megfelelője: *buzzy*). Azért, hogy ezt a jelenséget kiküszöböljék, számos továbbfejlesztett gerjesztési modellt javasoltak a szakirodalomban, melyeket különböző kategóriákba sorolhatunk a modell típusa és a gerjesztő jel szerint.



1. ábra. A HTS rendszerben lévő alap impulzus-zaj gerjesztés. Forrás: [8] alapján, módosítva.

A kevert gerjesztés és a STRAIGHT beszédkódoló használata [9] rendkívül jó minőségű HMM-alapú szintetizált beszédet eredményez, azonban ezek nehézkesen építhetők be valós idejű alkalmazásokba nagy számításigényük miatt. A kevert gerjesztés azon beszédhangok modellezésére különösen hasznos, amelyek nem egyértelműen zöngések vagy zöngétlenek, hanem ezek keverékeként jönnek létre (pl. zöngés réshangok).

A glottális (azaz gégében lévő) forrásjel leírása és paraméterekre bontása már régóta aktív kutatási terület. Cabral és társai a glottális forrás deriváltjának Liljencrants-Fant által kidolgozott (LF) akusztikus modelljét használják és glottális spektrális szétválasztást alkalmaznak [10]. Raitio és társai a korábban kidolgozott glottális inverz szűrés eljárást használják fel és integrálják a HTS rendszerbe,

melyet GlottHMM-nek neveznek [11]. Az egyetlen pulzust felhasználó technikát kiegészítik egy pulzus elem könyvtárral és elemkiválasztással; a legújabb eredmények szerint azonban a nagyméretű elemkönyvtár nem javítja a szintetizált beszéd minőségét [12]. Összességében a glottális forrást alkalmazó rendszerek jó minőségű zöngés beszédet tudnak létrehozni, de a zöngés-zöngétlen átmenetek kezelése nem teljesen megoldott és stabilitási problémák fordulhatnak elő.

Néhány módszer a harmonikus-zaj modell (*Harmonic Plus Noise Model*, HNM) alkalmazását javasolja a HTS környezetben és a paraméterek közé veszi a maximális zöngés frekvenciát [13] vagy zöngés vágási frekvenciát [14]. Ezen rendszerek előnye, hogy a spektrum felsőbb frekvencia sávjaiban sztochasztikus zajt alkalmazva csökkenthető a szintetizált beszéd zizegőssége.

Számos gerjesztési modell foglalkozik a beszéd maradékjelével. Ezen megoldások nagy előnye, hogy a maradékjel közvetlenül, automatikusan kinyerhető a beszédjelből lineáris predikció alapú inverz szűréssel. Az egyik ilyen modellben a maradékjel paraméterekkel történő leírására az amplitúdó spektrumot használják [14]. Egy másik gerjesztési modellben a hullámforma interpoláció valamint az idő- és frekvenciatartománybeli null-kitöltés a spektrális torzítást csökkenti [15]. Drugman és kollégái zöngeszinkron maradékjel kódkönyv építést alkalmaznak, majd PCA (*Principal Component Analysis*) eljárással tömörítik a kódkönyvet [16]. A módszer egyszerűsítéseként bevezetik a determinisztikus-sztochasztikus modellt (*Deterministic Plus Stochastic Model*, DSM), amely egy „sajátmaradékjel" újramintavételezésével állítja elő a maradékjel periódusokat [17]. A maradékjelen alapuló módszerek előnye, hogy könnyen kiegészíthetőek a normáltól eltérő zöngeminőségű beszéd modellezésére.

A statisztikai parametrikus beszédszintézis alapmódszereit és a legtöbb fenti gerjesztési modellt ideális beszédre dolgozták ki és optimalizálták. Azon beszélők esetén várhatóan nem eredményez jó minőséget, akiknél gyakran előfordul az ideálistól lényegesen eltérő zöngeképzés. Ennek egyik oka lehet az irreguláris fonáció.

2.2. Irreguláris zöngeképzés előfordulása és modellezése

Az emberi beszédben ideális (más néven reguláris vagy modális) zöngeképzés esetén a hangszalagok kváziperiodikusan rezegnek. A gégében azonban hosszabb-rövidebb időtartamra instabilitás léphet fel, ami a hangszalagok irreguláris rezgését okozza. Ez eltér a modális zöngeképzéstől, és irreguláris fonációnak, glottalizációnak, érdes zöngének vagy recsegő beszédnek nevezik. A kifejezés angol elnevezései: *irregular phonation, glottalization, creaky voice, vocal fry, laryngealization*. A jelenség a zöngeperiódusok hosszának és/vagy amplitúdójának hirte-

len megváltozásából adódik. Az irreguláris fonáció előfordul egészséges és patologikus beszélők esetén is, általában szakaszhatárokon (pl. mondat végén) vagy magánhangzó-magánhangzó kapcsolatban. Gyakran kíséri extrém alacsony alapfrekvencia és a glottális pulzusok gyors lecsökkenése [18]. Érzetileg recsegő, érdes jellegű beszédet jelent [5]. A glottalizáció a beszédhangok 15%-ában is előfordulhat egy-egy beszélő esetén, így egyáltalán nem elhanyagolható jelenség [19].

A 4. ábra egy példát mutat az irreguláris (a) és reguláris (e) fonáció hullámformájára (vízszintes vonal jelzi azt a szakaszt, ahol irreguláris a zöngeképzés). A glottalizáció a szokványos beszédfeldolgozó algoritmusok számára sokszor problémákat okoz (pl. automatikus $F0$ mérés és spektrális analízis). Az irreguláris fonáció megfelelő modellezése hozzájárulhat a természetes, expresszív és személyre szabott beszédszintetizátor rendszerek elkészítéséhez.

A szakirodalomban léteznek megoldások a reguláris-irreguláris osztályozásra [19, 20], a modális beszéd glottalizálttá transzformálására [5] és néhány kezdeti kísérletet végeztek statisztikai parametrikus beszédszintézissel is [21, 22, 23, 24]. Irreguláris-reguláris transzformációra nem találtunk eljárást a szakirodalomban. A beszédszintézisben Silén és társai által bemutatott módszer lényege, hogy robusztus $F0$ mérést alkalmaz megbízható zöngésség detekcióval, ezáltal eltüntetve a glottalizált beszédrészleteket a szintetizált beszédből [21]. Így viszont a beszélőre jellemző irreguláris fonáció teljesen elveszik a beszédszintézis kimenetéből. Drugman és társai a DSM modell továbbfejlesztésével analízis-szintézis kísérletekben bemutatják, hogy a maradékjel periódusokban előforduló másodlagos impulzusok jelenléte megfelelően modellezi az irreguláris beszédet [22]. Ezután megvizsgálják, hogy a HTS rendszer mely környezetfüggő címkéi lehetnek hasznosak a glottalizáció előfordulásának előrejelzésére és új paraméterfolyamokat is hozzáadnak a rendszerhez, amelyek segítik az automatikus döntést az irreguláris zönge helyéről [23]. Raitio és társai egyesítik a fenti módszereket és bemutatnak egy irreguláris zönge előrejelzésére és szintézisére alkalmas rendszert a DSM és GlottHMM modellek kiegészítéseként [24]. Eredményeik szerint a glottalizált minták használata kis mértékben érdesebbé tette a szintetizált beszédet, míg nem javította az alaprendszer természetességét.

2.3. Szubglottális rezonanciák hatása a beszédre

Beszédhangjaink akusztikai minőségét nem csak a gége és a felette lévő szervek határozzák meg, hanem a gége alatti (szubglottális) légzőszervek bizonyos tulajdonságai (pl. tüdő térfogata, légcső hossza) is befolyásolják azt. A forrás-szűrő modell tehát a valóságban nem tökéletes modellje a beszédképzésnek, mivel nemlineáris csatolás jöhet létre a forrás és a szűrő között. Az alsó légúti rendszer

rezonanciái (szubglottális rezonanciák, SGR) a formánsokhoz hasonlóan alakítják a zöngés hangok spektrumát, az SGR-ek frekvenciájának környezete azonban akusztikai szempontból előnytelen. Emiatt azt feltételezik, hogy a beszéd képzése során próbáljuk elkerülni azokat az artikulációs helyzeteket, amikor a formánsok és szubglottális rezonanciák között interakció léphetne fel, így a formánsok is elkerülik az SGR értékeket. Mivel az alsó légúti szervek közül a légcső és a hörgők fiziológiai méretei viszonylag keveset változnak a beszéd során, a rezonanciafrekvenciák közel állandóak egy-egy ember beszédében. Az első három szubglottális rezonancia tipikus értéke 600, 1500 és 2300 Hz körül mérhető [6].

Az utóbbi években több nyelvre (amerikai angol [7], spanyol, német és koreai) megmutatták, hogy az alsó légutak rezonanciái a magánhangzókat és a mássalhangzókat a frekvenciaszerkezetük szerint diszkrét csoportokra bontják, melyek jellegzetes kategóriáknak feleltethetőek meg (fonológiai megkülönböztető jegyek, [6]). Ezen kategóriákat már számos elméleti megközelítés segítségével próbálták magyarázni, melyek közül az egyik legsikeresebb a kvantális elmélet (QT, Quantal Theory) [25]. A kvantális elmélet azon alapul, hogy a beszédhangokban mérhető akusztikai paraméterek és a beszélő által változtatott artikulációs helyzetek jellegzetes nem-monoton módon változnak, azaz az artikulációs tér egyes részeiben lévő kis változások nagy akusztikus változáshoz vezetnek, míg más, nagyobb artikulációs változtatások csak kisebb akusztikus változással járnak [6]. Ennek egyik jó példája az elöl és hátul képzett magánhangzók esete: a két magánhangzó csoportot a nyelv vízszintes mozgása különbözteti meg. A kvantális elmélet szubglottális rezonanciákra vonatkozó kiegészítése szerint az $Sg2$ egy természetes elválasztó az elöl és hátul képzett magánhangzók között amerikai angolban [7]. Az első szubglottális rezonancia ($Sg1$) hatása általában kevésbé erős, mégis azt vették észre, hogy az első formáns ($F1$) tekintetében az $Sg1$ elválasztó szerepet játszik az alsó és nem-alsó magánhangzók között. A harmadik szubglottális rezonancia ($Sg3$) sokszor az elöl képzett feszes és laza magánhangzók között helyezkedik el amerikai angolban [7].

Az eddigi eredmények szerint a szubglottális rezonanciák a formánsmenetekben a folytonosság megszakadását okozhatják, észrevehetőek a beszédpercepció számára, valamint Wang és társai kutatásai szerint hasznosak lehetnek az automatikus beszélő normalizálásban [26]. Eddig azonban csak néhány nyelvre vizsgálták a magánhangzó formánsok és SGR-ek kapcsolatát. A szubglottális rezonanciák beszédhangokra kifejtett szerepével kapcsolatban magyar nyelvre korábban nem történt kutatás.

3. Kutatási célkitűzések

Kutatásaimmal a rejtett Markov-modell alapú gépi szövegfelolvasók természetességének növeléséhez és a beszédképzés forrás-szűrő modelljének pontosításához kívánok hozzájárulni. Konkrét céljaim a kutatás során:

1) a statisztikai parametrikus beszédszintézisben a gépi beszéd természetességének növelése,
2) irreguláris zöngeképzés elemzése és ennek javítása, rekedtes beszéd hangzásának kellemesebbé tételére,
3) irreguláris beszédmodellek létrehozása beszédszintézisben, amelyekkel expresszív és személyre szabható gépi szövegfelolvasó rendszerek készíthetőek,
4) az emberi beszédképzésben a forrás-szűrő közti kölcsönhatás pontosabb megismerése, különös tekintettel a szubglottális rendszer hatására.

Ezeket a kutatási célokat azért választottam, mert számos kihívást tartalmaznak és a kutatásommal hozzá tudok járulni az ember-gép kapcsolat természetesebbé tételéhez. Munkám során a kísérleteket magyar nyelvű beszédkorpuszokon végeztem, de az eredmények nagy része könnyen alkalmazható más nyelvekre is. Az I. téziscsoportban az 1) és 2) kutatási célokkal, a II. téziscsoportban az 1) és 3) célokkal foglalkozom, míg a III. téziscsoportban a 4) kutatási célt teljesítem.

4. Módszertan

Kutatásom során a létrehozott módszerek eredményességét kísérleti úton vizsgáltam.

A beszéd analízisével, szintézisével és az irreguláris zöngeképzéssel kapcsolatos kísérleteket (I. és II. téziscsoportok) a PPBA adatbázisból kiválasztott 5 magyar anyanyelvű beszélőn végeztem [27]. Négy férfitól (FF1, FF2, FF3 és FF4) és egy nőtől (NO3) származó, professzionális körülmények között rögzített, kb. 2-2 órányi hangfelvételt használtam fel. A módszereimet magyar mintákon teszteltem és validáltam, de az itt alkalmazott eljárások nyelvfüggetlenek és várhatóan más nyelvre is hasonló módon alkalmazhatóak.

A szubglottális rezonanciák vizsgálatához (III. téziscsoport) a beszéd és szubglottális felvételek egy részét a kutatás során rögzítettük magyar anyanyelvű beszélőkkel[2]. Részben négy beszélő logatom felvételein [C4] (két férfi: Log_FF1, Log_FF2 és két nő: Log_NO1, Log_NO2), részben a BEA adatbázis [28] hat beszélőjétől származó spontán beszéd felvételeken és ugyanezen beszélők logatom felvételein [J4] végeztem az elemzéseket (öt férfi: Spo_FF1 – Spo_FF5 és egy nő:

[2] A többes szám a kutatásban részt vevő többi személyre utal: Bárkányi Zsuzsanna, Gráczi Tekla Etelka, Bőhm Tamás, Beke András és Steven M. Lulich. A felvételek készítését, a manuális méréseket és a kézi javításokat közösen végeztük.

Spo_NO1). A felvételeket csendes szobában végeztük, a szubglottális jelet gyorsulásmérő eszközzel rögzítettük, melyet a nyakon a gégénél lévő pajzsporchoz szorítottunk. A szöveges és fonetikus átírást valamint a hanghatárok címkézését a kutatás során készítettük el automatikus eszközökkel és manuális javítással.

Kutatásaim során a következő eszközöket és szoftvereket használtam fel:

BME-TMIT kényszerített felismerő: hanghatárok automatikus címkézése,
GLOAT / SEDREAMS: beszédjel felbontása zöngeszinkron periódusokra, http://tcts.fpms.ac.be/~drugman/Toolbox/
HTS: paraméterek tanítása HMM-ek segítségével [2], http://hts.sp.nitech.ac.jp/
HTS-HUN: a HTS rendszer magyar változata [8]
Matlab: beszédjel analízise és szintézise, ROC elemzés, t-teszt
Praat: alapfrekvencia mérése; formánsok mérése; beszédjel vizuális elemzése, http://www.fon.hum.uva.nl/praat/
SoX: beszédjel aluláteresztő szűrése és újramintavételezése, http://sox.sourceforge.net/
SPTK: spektrális elemzés, inverz szűrés és digitális szűrés, http://sp-tk.sourceforge.net/
Voice_Analysis_Toolkit / creak_detect: irreguláris zönge detektor [20], https://github.com/jckane/Voice_Analysis_Toolkit
VoiceSauce: beszédjel akusztikai paramétereinek korrekciója; Harmonics-to-Noise Ratio számítása, http://www.ee.ucla.edu/~spapl/voicesauce/
Wavesurfer: beszédjel és gyorsulásmérő jel vizuális elemzése és akusztikai paraméterek mérése, http://www.speech.kth.se/wavesurfer/
Weka: döntési fák megvalósítása, http://www.cs.waikato.ac.nz/ml/weka/.

A transzformációs eljárások és szintézis módszerek eredményességét percepciós (meghallgatásos) kísérletekkel is vizsgáltam. A kísérletek készítése során a szakirodalomban javasolt teszt típusokból indultam ki. A meghallgatásos tesztekben a tesztelők az egyes hangminták meghallgatása után 1–5 skálás MOS (*Mean Opinion Score*), illetve minta párok esetén 1–3 vagy 1–5 skálás CMOS (*Comparative Mean Opinion Score*) jellegű kérdésekre válaszoltak. Az egyes kísérletek körülményei és részletei a disszertációban olvashatóak.

A statisztikai elemzések során egymintás t-tesztet, párosított mintás t-tesztet és Tukey-HSD post-hoc teszttel kiegészített egytényezős ANOVA analízist alkalmaztam a Matlab és SPSS programokkal. Az elemzések során kétoldalas $p < 0{,}05$ szignifikancia szint alatt (95% konfidencia szint felett) vetem el a nullhipotézist.

A tézisfüzetben használt rövidítéseket és jelöléseket a füzet végén összesítve felsorolom.

5. Új eredmények

I. téziscsoport: Új, MGC maradékjel kódkönyv alapú gerjesztési modell kidolgozása és felhasználása irreguláris zöngeképzés javítására

A szakirodalomban számos beszéd analízis-szintézis módszerről olvashatunk, melyeknek célja eredetileg a beszéd paraméterekre bontása és kódolása volt azért, hogy a távközlési csatornán minél kisebb sávszélesség mellett lehessen átvinni jól érthető beszédet. Emellett napjainkban a beszédfeldolgozás területén egyre fontosabb, hogy a beszédjel olyan parametrikus felbontását találjuk meg, amely különböző transzformációkra alkalmazható és gépi tanuló rendszerben is felhasználható. Kezdeti kísérleteink[3] szerint a ma elérhető legjobb beszédkódoló eljárások (pl. CELP, azaz *Code-Excited Linear Prediction* jellegű kódolók) nem alkalmasak a gépi tanulórendszerbe történő integrálásra (pl. a CELP kódoló kódkönyv indexe ugráló értékeket tartalmaz, ami nem modellezhető egyszerűen HMM-ekkel). A 2.1. fejezetben ismertetett gerjesztési modellek közül az egyszerűbbek (pl. impulzus-zaj modell) zizegős beszédet eredményeznek, a bonyolultabbak (pl. kevert gerjesztés) ugyan jobb minőségű beszéd szintetizálható, de sokszor nehezen használhatóak fel valós idejű alkalmazásokban nagy számításigényük miatt. A két véglet között olyan gerjesztési modell elkészítését céloztuk meg, melynek minősége megfelelő, és várhatóan használható korlátozott erőforrású eszközben is.

A téziscsoport bemutat egy beszédet paraméterekre bontó, maradékjelen alapuló, nyelvfüggetlen gerjesztési modellt, amely beszéd analízis-szintézisére alkalmas és a paraméterei integrálhatóak a rejtett Markov-modell alapú gépi tanításba. A korábbi eljárások közül vannak ehhez hasonló gerjesztési modellek. A DSM eljárás is maradékjel kódkönyv alapú, azonban ez nem alkalmaz összefűzési költséget az elemkiválasztás során [16]. A GlottHMM rendszerben alkalmaznak ugyan célköltséget és összefűzési költséget is, de ez glottális forrásjel szintjén történik [12]. Ez alapján az I.1. tézisben javasolt modell lényeges pontokban különbözik az általam ismert korábbi rendszerektől. Emellett új típusú, korábban nem használt paramétereket vezetek be a maradékjel leírására. A további tézispontokban ismertetem a modell alkalmazását az irreguláris zöngével képzett természetes beszéd érzeti érdességének csökkentésére.

I.1. tézis: [C3] Új, maradékjel kódkönyv elemkiválasztás alapú nyelvfüggetlen gerjesztési modellt dolgoztam ki, amely a beszédjel paraméterekre bontására (analízis) és visszaállítására (szintézis) alkalmas. A módszerben beszéd maradékjel halmaz alapján zöngeszinkron periódusokból álló kódkönyv készül, melyek-

[3] A továbbiakban többes szám első személyt használok a könnyebb olvashatóság érdekében. Saját eredményeimet a tézisekben összegzem.

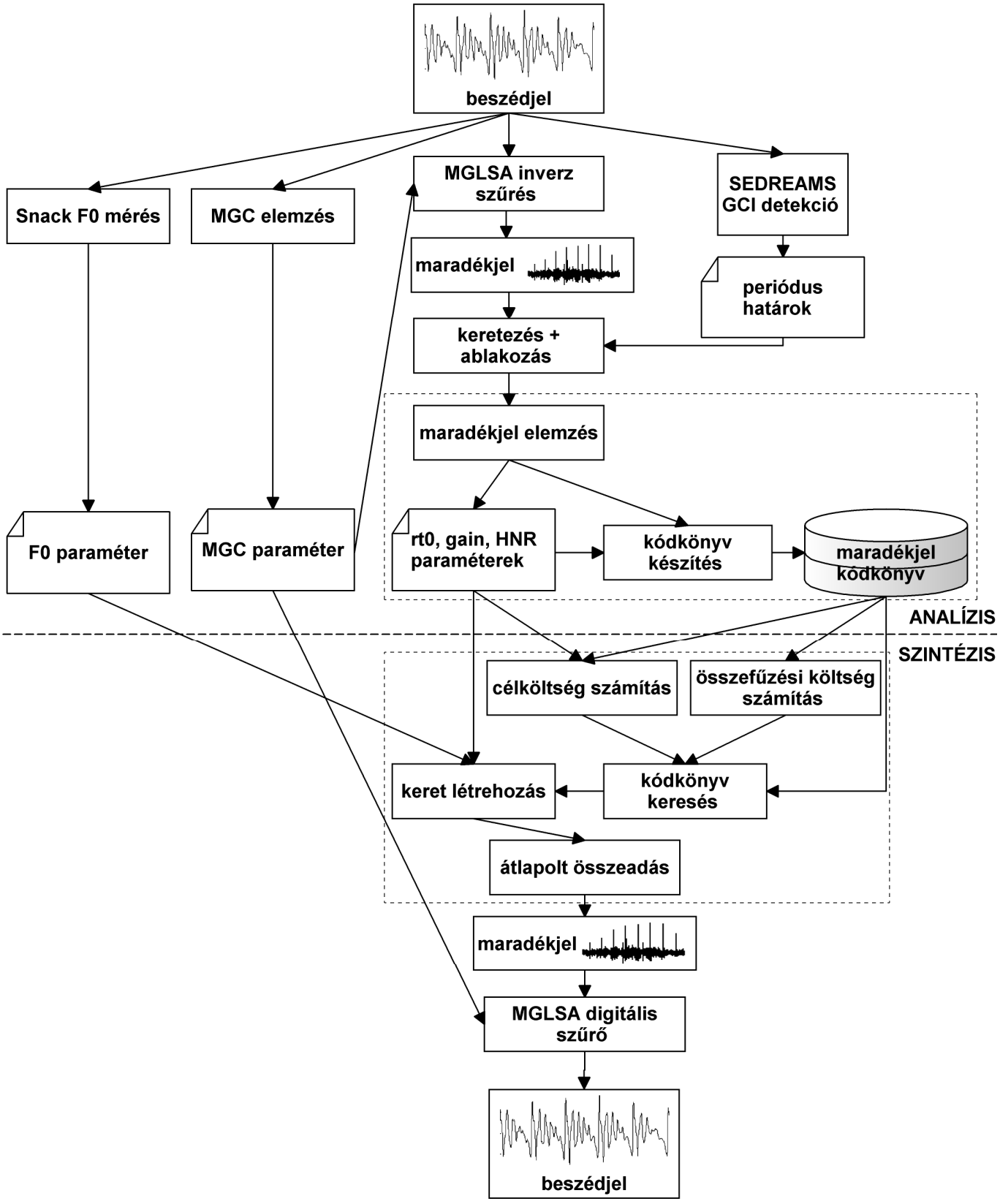
ből szintézis során automatikus elemkiválasztás határozza meg az összeillesztendő elemeket, célköltséget és összefűzési költséget felhasználva.

Az analízis lépései a 2. ábra szaggatott vonal feletti részén láthatóak. Az analízis módszer bemenete beszéd hullámforma, amelyet 7,6 kHz-es aluláteresztő szűrés után 16 kHz mintavételezéssel és 16 bites lineáris PCM kvantálással tárolunk. A módszer először egy zöngeszinkron maradékjel periódusokból álló kódkönyvet épít, majd elvégzi a maradékjel elemzését. A beszéd alapfrekvenciáját 25 ms kerethosszal és 5 ms eltolással mérjük a Snack $F0$-detektáló algoritmusával. A következő lépésben spektrális elemzést végzünk MGC (*Mel-Generalized Cepstrum*, magyarul *Mel-Általánosított Kepsztrum*) módszerrel. Az elemzéshez 34-ed rendű MGC analízist alkalmazunk $\alpha = 0{,}42$ és $\gamma = -1/3$ paraméterekkel. A maradékjelet, vagyis a beszéd gerjesztését MGLSA (*Mel-Generalized Log Spectral Approximation*) inverz szűréssel számoljuk. Ezután az SEDREAMS (*Speech Event Detection using the Residual Excitation And a Mean-based Signal*) zöngeperiódus-meghatározó algoritmust alkalmazzuk a zöngés maradékjel periódusainak szétválasztásához.

Az analízis további lépéseit a maradékjelen végezzük el 50 ms keretméret és 5 ms eltolás értékekkel. A zöngés szakaszokból zöngeszinkron, két periódus hosszú, Hann-ablakozott maradékjel periódusokat vágunk ki, melyekből egy kódkönyv készül. A kódkönyv elemek leírására a következő paramétereket használjuk:

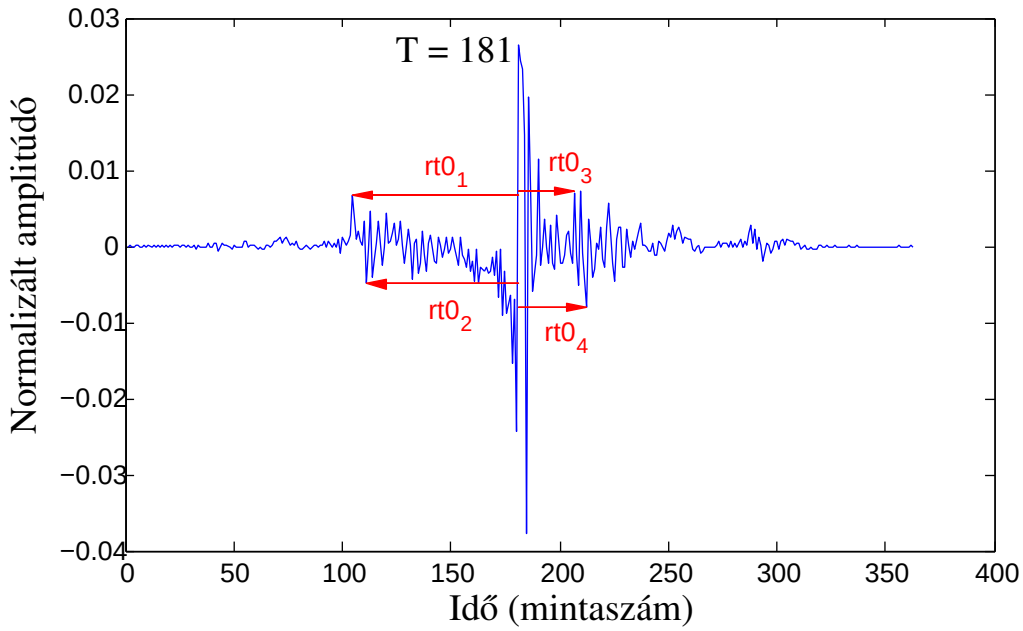
F0: az elem alapfrekvenciája,
gain: az elem energiája:
$gain_i = \sqrt{\sum_{j=0}^{N} r_j^2}$, ahol r_j az $i.$ ablakozott elem $j.$ mintája,
rt0: az ablakozott elemben a kiugró csúcsok pozíciója (példa: 3. ábra),
HNR: az elem harmonikus-zaj aránya (*Harmonics-to-Noise Ratio*) kepsztrális harmonikus alapon [29].

Minden zöngés kerethez eltárolunk egy kódkönyv elemet az ablakozott jellel és a fenti paraméterekkel együtt. Az $rt0$ paraméter célja az ablakozott maradékjel kódkönyv elemben lévő jelentős csúcsok leírása, melyre a 3. ábra mutat példát. Korábban ilyen paramétert használó megoldást egyik módszer sem alkalmazott a maradékjel leírására. A paraméter számítási módja részletesen a disszertációban található. A maradékjel kódkönyv készítése során a hasonló, egymáshoz várhatóan illeszkedő elemeket összefűzési költség felhasználásával számítjuk: az elemeket normalizáljuk majd RMSE (*Root Mean Squared Error*) távolságot számítunk. A beszédjel analízise során a fenti paramétereket nyerjük ki minden zöngés keretből (azaz ha $F0 > 0$). Zöngétlen keret esetén ($F0 = 0$) csak a $gain$ értéket számoljuk.



2. ábra. Beszédjel analízise (szaggatott vonal felett) és szintézise (szaggatott vonal alatt) az MGC maradékjel kódkönyv alapú módszerrel. Négyzetek jelölik az eljárásokat és hullámformákat; a behajtott sarkú négyzetek a paramétereket jelzik.

A szaggatott vonalú téglalapok mutatják az általam hozzáadott új eljárásokat.



3. ábra. Az $rt0$ paraméter számítása egy ablakozott maradékjel kódkönyv elemre. Az $rt0_i$ érték a kiugró csúcsok mintában mért távolságát adja meg az elemben lévő impulzushoz ($T = 181$) képest. Az ábrán lévő értékek: $rt0_3 < rt0_4 < rt0_2 < rt0_1$.

A szintézis lépéseit a 2. ábra szaggatott vonal alatti része mutatja be. A szintézis bemenete az analízis eredményeként kapott paraméterek ($F0$, $gain$, $rt0$, HNR és MGC) illetve a zöngeszinkron maradékjelek kódkönyve. A visszaállítás során először a maradékjelet állítjuk elő keretenként. Amennyiben a keret zöngés ($F0 > 0$), az $F0$, $rt0$ és HNR paraméterek alapján egy megfelelő, hozzá tartozó elemet keresünk a kódkönyvből. Kézzel beállított súlyozású célköltséget és összefűzési költséget alkalmazunk az elemkiválasztásos beszédszintézishez hasonlóan [4]. A célköltség az aktuális keret és a kódkönyv elemeinek paraméterei közötti négyzetes különbség. Az összefűzési költséget a kódkönyv elemek normalizált változatának átlagos négyzetes különbségeként (RMSE távolság) számítjuk. A legmegfelelőbb kódkönyv elem hosszát a cél $F0$-nak megfelelően beállítjuk törléssel vagy nullák hozzáadásával. Amennyiben a keret zöngétlen ($F0 = 0$), fehér zajt használunk gerjesztésként. Ezután a maradékjelet a Hann-ablakozott periódusok zöngeszinkron átlapolt összeadásával és a zöngétlen részek összefűzésével kapjuk. Végül a keretek energiáját a $gain$ paraméter alapján beállítjuk, majd a szintetizált beszédet előállítjuk MGLSA szűréssel az MGC paramétereket felhasználva.

I.2. tézis: [C1, C5] Nyelvfüggetlen eljárást dolgoztam ki irreguláris zöngével képzett beszéd regulárissá alakítására az I.1. tézisben kidolgozott modell felhasználásával. Percepciós teszttel kimutattam magyar mintákon, hogy az irreguláris-

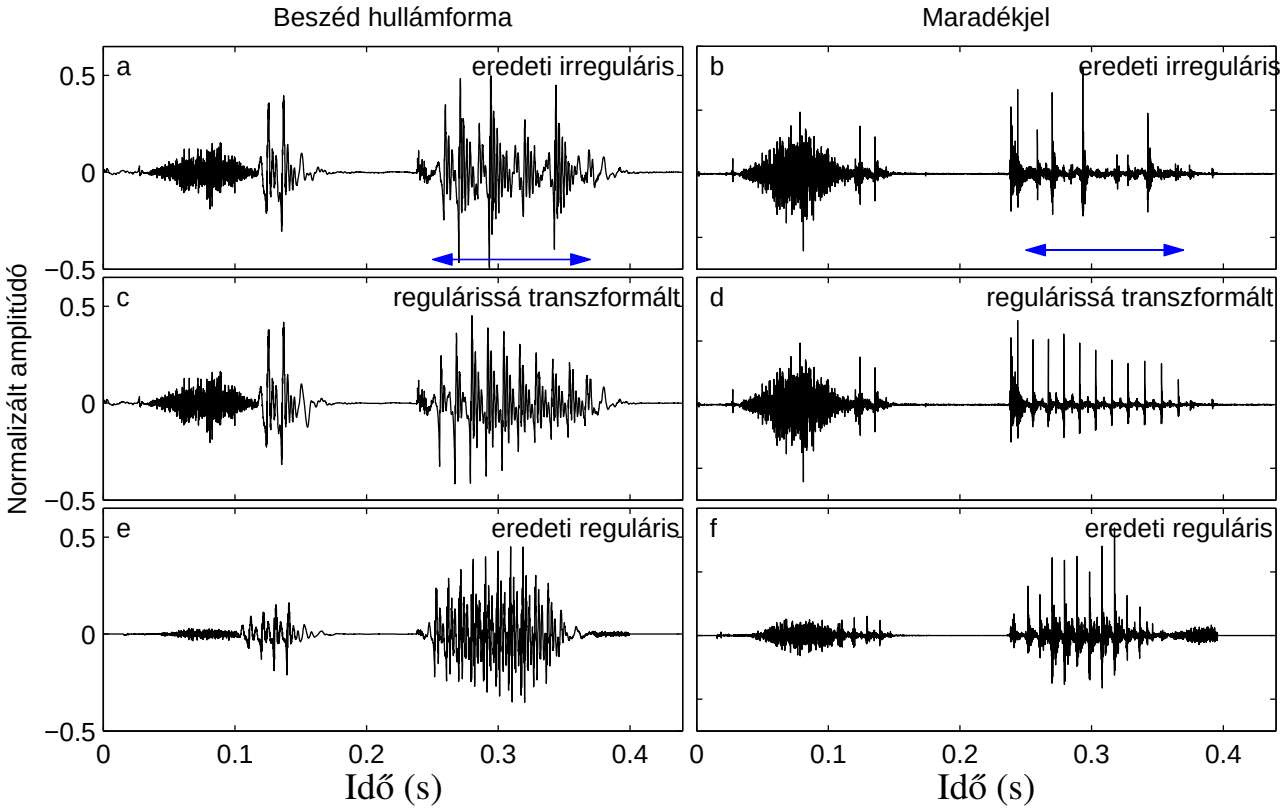
reguláris transzformáció után a beszéd szignifikánsan kevésbé érdes, mint az eredeti irreguláris beszéd.

A kidolgozott eljárás az I.1. tézis analízis-szintézis módszerét egészíti ki egy olyan transzformációs eljárássá, amely alkalmas a glottalizált beszéd modálissá alakítására, tehát az irreguláris zöngeképzés javítására. Az analízis hasonlóan történik, mint az I.1. tézisben, azzal a különbséggel, hogy a kódkönyvet csak modális maradékjel szakaszokból építjük, az irreguláris zöngével képzett részeket kihagyva. Az analízis után a paramétereket módosítjuk, majd az I.1. tézis szintézisével visszaállítjuk a javított beszédjelet.

A transzformáció során az eredeti beszéd maradékjelnek azon szakaszait módosítjuk, amelyet irreguláris zönge címkék jeleznek, míg a modális zöngés és zöngétlen maradékjel részeket változatlanul hagyjuk. Az analízis eredményeként kapott $F0$ értékeket interpoláljuk, míg a $gain$ és MGC értékeket simítjuk az irreguláris szakaszokon. A glottalizáció hibákat okozhat az $F0$ detekcióban: a hirtelen alapfrekvencia és amplitúdó változás miatt előfordulhat, hogy egy eredetileg zöngés keretet zöngétlennek jelöl a detektor, vagy az eredeti érték felét méri. Emiatt a mért $F0$-menetet interpoláljuk azokban a zöngés szakaszokban, ahol az algoritmus nem detektált zöngét. A kísérletek során minden $F0$-menetet kézzel ellenőriztünk és javítottunk, emiatt a módszer félautomatikus működésű. Az irreguláris fonáció kis perturbációkat okoz a keretenkénti $gain$ és MGC értékekben az irreguláris zöngeperiódusok amplitúdójának hirtelen változása miatt. Emiatt 5-pontos simítást végeztünk ezeken a paramétereken, amely tapasztalataink szerint megfelelőnek bizonyult a perturbációk eltüntetésére. A szintézis további lépései megegyeznek az I.1. tézisben ismertetett lépésekkel.

Az irreguláris-reguláris transzformáció eredményére láthatunk egy példát a 4. ábrán. Az ábrán észrevehető, hogy a „regulárissá transzformált” (c és d) és az „eredeti reguláris” (e és f) változatoknak hasonló zöngeperiódusai vannak, míg az „eredeti irreguláris” (a és b) jel ettől lényegesen eltérő és periódusonkénti amplitúdó ingadozást tartalmaz.

Az irreguláris-reguláris transzformáció működését a PPBA adatbázis négy beszélőjének (3 férfi: FF1, FF3 és FF4 és egy nő: NO3) hanganyagán teszteltük [27]. Kiválasztottunk 4-4 szót, amelyek reguláris és irreguláris formában is előfordultak az adatbázisban. Ezután az irreguláris változatot transzformáltuk a fenti módszerrel. A szavak 3-3 változatát (eredeti irreguláris, regulárissá transzformált és eredeti reguláris) meghallgatásos tesztben hasonlítottuk össze, melyet 9 tesztelő végzett el. Az eredményeket párosított mintás t-teszttel összehasonlítva megállapítottuk, hogy a transzformáció mind a négy beszélő esetén szignifikánsan csökkentette az érzeti érdességet ($p < 0{,}05$) az eredeti irreguláris változathoz képest.



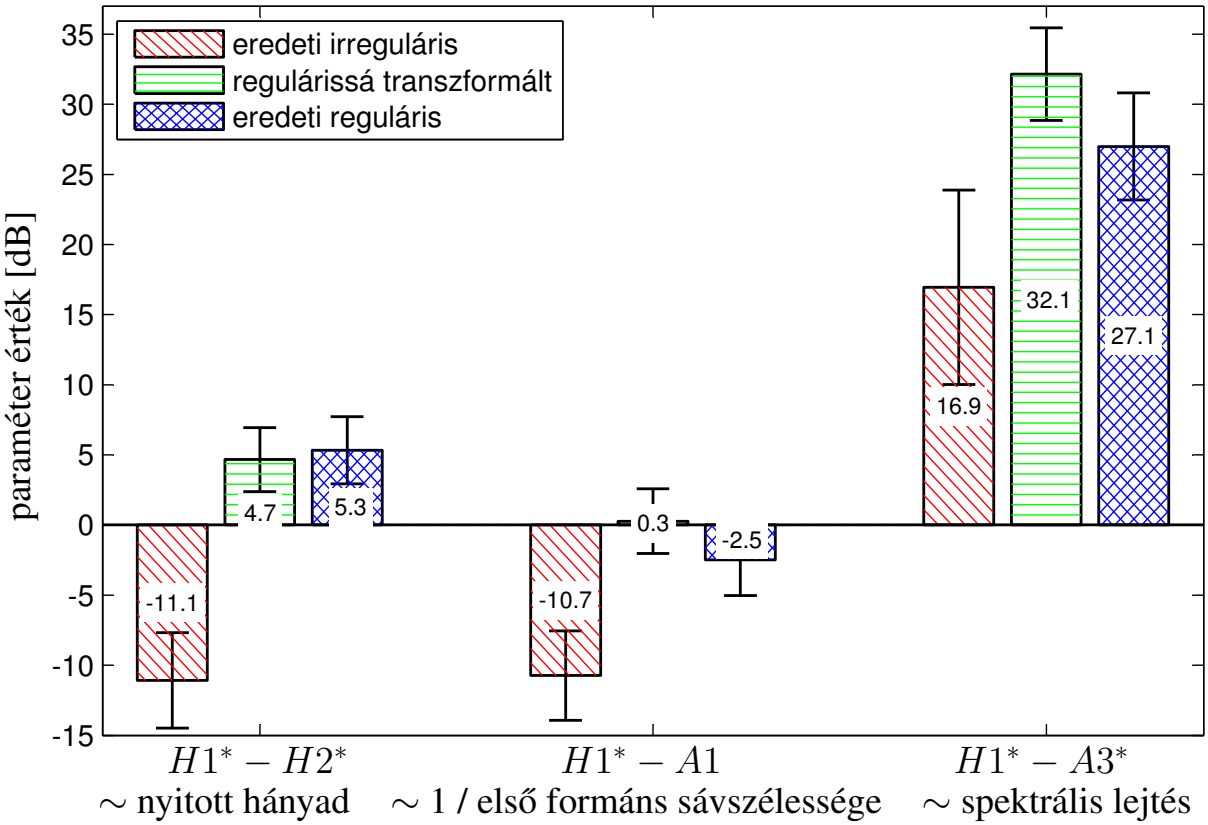
4. ábra. A kiejtett és transzformált „cipő” szó beszéd hullámformái és maradékjelei FF3 beszélőtől: a) beszédjel és b) maradékjel eredeti irreguláris záró magánhangzóval (nyíl jelöli az irreguláris zöngét), c) beszédjel és d) maradékjel regulárissá transzformált záró magánhangzóval, e) beszédjel és f) maradékjel eredeti reguláris záró magánhangzóval (a szó másik realizációja).

I.3. tézis: [C1, C5] Kísérleti úton igazoltam magyar mintákon, hogy az I.2. tézis eljárása a beszéd több releváns akusztikai paraméterét (nyitott hányad, első formáns sávszélessége, spektrális lejtés) az irreguláris-reguláris transzformáció során a reguláris zöngeképzésre jellemző értékek irányába módosítja.

Az I.2. tézisben meghallgatásos teszthez kiválasztott beszédmintákon (eredeti irreguláris, regulárissá transzformált, eredeti reguláris) akusztikus elemzést is végeztünk. A szakirodalomból kiválasztottunk három olyan akusztikai jegyet, amelyeket korábban irreguláris és reguláris beszéd megkülönböztetésére használtak [30, 5]. Ezek alapján irreguláris zöngeképzés esetén a nyitott hányad (*open quotient*, OQ) alacsonyabb; az első formáns sávszélessége (*first formant bandwidth*, $B1$) nagyobb; a spektrum lejtése (*spectral tilt*, TL) meredekebb, mint reguláris beszédben.

A transzformáció hatását az OQ, $B1$, TL akusztikai jellemzőkre mérésekkel vizsgáltuk. A méréseket frekvenciatartományban végeztük [31] a Wavesurfer programban vizuálisan leolvasva, a paraméterek korrekciójával [32]: az OQ megfeleltethető az első és második harmonikus dB-ben mért különbségének ($H1^* -$

$H2^*$), $B1$ reciproka arányos $H1$ és az első formáns amplitúdójának különbségével ($H1^* - A1$), míg a TL korrelál $H1$ és a harmadik formáns amplitúdójának különbségével ($H1^* - A3^*$). A mérési eljárás pontos részletei a disszertációban olvashatóak.



5. ábra. Az irreguláris-reguláris transzformációval módosított szavak akusztikus elemzésének eredménye. A függőleges fekete vonalak a 95%-os konfidenciaintervallumot jelölik.

A három mért akusztikai paramétert a három beszédminta típuson az 5. ábra mutatja be. ANOVA elemzést és Tukey-HSD post-hoc tesztet végezve megállapítottuk, hogy a $H1^* - H2^*$ megközelítőleg azonos az eredeti reguláris és a transzformált beszédrészleteken ($p = 0{,}938$, n.s. különbség), míg szignifikánsan különböző az eredeti irreguláris mintákhoz képest ($p < 0{,}0005$). A nyitott hányad szempontjából a transzformált változatok tehát közel vannak a modális beszédhez. Az irreguláris zöngével képzett szavak $H1^* - A1$ és $H1^* - A3^*$ különbségei szintén szignifikánsan különbözőek az eredeti reguláris és regulárissá transzformált változatokhoz képest ($p < 0{,}0005$ és $p < 0{,}05$), de az eredeti reguláris és a transzformált változatokban közel megegyeznek ($p = 0{,}336$ és $p = 0{,}321$, n.s. különbség). Eszerint a transzformált minták közel vannak az eredeti modális felvételekhez $B1$ és TL tekintetében is. A transzformációs eljárás a vizsgált akusz-

tikai jegyek szempontjából tehát a reguláris beszédre jellemző értékek irányába módosítja az irreguláris beszédjelet.

II. téziscsoport: Új gerjesztési modell illesztése gépi szövegfelolvasóhoz és felhasználása irreguláris beszéd szintézisére

Az irodalmi áttekintés bemutatott számos gerjesztési modellt, amelyeket statisztikai parametrikus beszédszintézisben alkalmaznak. A módszerek egy része kevert gerjesztést használ, más eljárások a glottális forrásjelet próbálják modellezni, bizonyos kísérletekben a harmonikus-zaj modellt fejlesztik tovább, és jónéhány esetben beszéd maradékjel alapú modellt alkalmaznak.

Ebben a téziscsoportban az I.1. tézis maradékjel alapú gerjesztési modelljét statisztikai parametrikus beszédszintézisbe illesztem. A javasolt rendszert a HTS szabadon hozzáférhető változatával, az impulzus-zaj gerjesztéssel hasonlítom össze. Ezután a javasolt rendszert kiegészítem két alternatív irreguláris zönge modellel.

II.1. tézis: [J2] Rejtett Markov-modell alapú gépi szövegfelolvasó rendszerhez illesztettem az I.1. tézisben ismertetett nyelvfüggetlen gerjesztési modellt. Percepciós teszttel igazoltam magyar mintákon, hogy a módszerrel előállítható beszéd szignifikánsan jobb minőségű az impulzus-zaj gerjesztésű gépi szövegfelolvasóhoz képest.

Az I.1. tézis analízis lépésénél leírt paramétereket ($F0$, $gain$, $rt0$, HNR és MGC) kiszámítjuk a tanító beszédadatbázis mondatainak minden 50 ms-os keretére, 5 ms-os eltolással. A paraméterek derivált és második derivált értékeit is eltároljuk a paraméterfolyamban. Ezután a tanításhoz a HTS-HUN rendszerhez [8] illesztettük a paramétereket. A beszédadatbázis fonetikus átirataiból környezetfüggő címkézés készül. A változó dimenziójú $log(F0)$, $log(rt0)$ és $log(HNR)$ paramétereket MSD-HMM-mel modellezzük (az $F0$-hoz hasonlóan az $rt0$ és HNR paraméterek valós értékűek a zöngés keretekre, de nem értelmezettek zöngétlen esetben). A logaritmus értékek használata a kísérletek során jobb eredményre vezetett. A többi paramétert ($log(gain)$ és MGC) hagyományos HMM-ek modellezik. Az időtartamok modellezéséhez minden fonémára beszédállapot időtartam eloszlásokat számít a rendszer. A fonéma-függő állapot időtartamokat Gauss eloszlással modellezzük. A környezetfüggő címkézés és az alkalmazott döntési fák csökkentik az összes lehetséges hangkörnyezet kombinációját. Az egyes paraméterfolyamokat külön döntési fákkal kezeljük.

A szintézis az I.1. tézisben leírthoz hasonlóan megy végbe néhány kiegészítéssel. A gépi tanulás eredményeként kapott $F0$, $gain$, $rt0$ és HNR paraméterek és

a maradékjel kódkönyv segítségével előállítjuk a maradékjelet. Ezután 6 kHz-es aluláteresztő szűrést végzünk, és a 6 kHz feletti frekvencia tartományban fehér zajt használunk a HNM alapú modellekhez hasonlóan. Erre a lépésre azért van szükség, mert lényegesen csökkenti a zöngés hangoknál előforduló zizegősséget. Végül a beszédet az MGC paraméterek segítségével szintetizáljuk MGLSA szűrővel. Az új rendszert HTS-CDBK-nak nevezzük.

A PPBA adatbázis FF2 férfi beszélőjének hanganyagával végeztünk beszédszintézis kísérleteket. Ehhez a teljes, 137 percnyi (1938 mondat) beszédfelvételt és a hozzá tartozó címkézést használtuk fel beszélő függő tanítás keretében. Az eredetileg 44,1 kHz-en tárolt mintákat újramintavételeztük 16 kHz-en 7,6 kHz-es aluláteresztő szűrés után. Alaprendszerként a HTS-HUN egyszerű impulzus-zaj gerjesztésű változatát (HTS-PN) használtuk. Az FF2 beszélő maradékjelei alapján 6 500 elemből álló kódkönyvet készítettünk a HTS-CDBK rendszerben. Mindkét rendszerrel 130-130 olyan mondatot szintetizáltunk, amely nem fordult elő a tanító adatbázisban. 20-20 mondatot kiválasztottunk egy meghallgatásos teszthez, amelyben a minták minőségét értékelte 15 tesztelő páros összehasonlítás keretében. A statisztikai elemzés szerint a HTS-CDBK rendszert szignifikánsan ($p < 0{,}0005$) jobb minőségűnek értékelték a HTS-PN rendszerhez képest.

A statisztikai parametrikus beszédszintézis és a legtöbb ebben használt gerjesztési modell (így a II.1. tézisben ismertetett módszer is) ideális beszéd esetén működik megfelelően, és számos hibát eredményez nem-modális zöngeképzés, például irreguláris fonáció esetén. A glottalizált beszédszakaszokon (általában a mondatok utolsó szótagjában) az $F0$-mérő algoritmus nem megfelelően méri az $F0$-t és zöngétlennek ítéli a keretet. Ezt a mintázatot a gépi tanulás is megtanulja, és az irreguláris fonációt a HTS-CDBK rendszer a zöngétlen beszédhez hasonlóan modellezi. Ez kellemetlen, rossz minőségű hangzást okoz, és nem megfelelő modellje a glottalizációnak. A továbbiakban a II.1. tézisben ismertetett HTS-CDBK rendszert használjuk alaprendszernek és ezt egészítjük ki irreguláris zönge modellekkel.

II.2. tézis: [C2, J1] Kidolgoztam egy nyelvfüggetlen szabály alapú irreguláris zöngeképzés modellt és illesztettem ezt a II.1. tézisben ismertetett gépi szövegfelolvasóhoz. A modell alapfrekvencia felezést, maradékjel periódus amplitúdó skálázást és spektrális torzítást alkalmaz. Percepciós teszttel igazoltam magyar mintákon, hogy a kiegészített rendszerrel szintetizált beszéd szignifikánsan preferáltabb és jobban emlékeztet az eredeti beszélőre, mint a II.1. tézis rendszere.

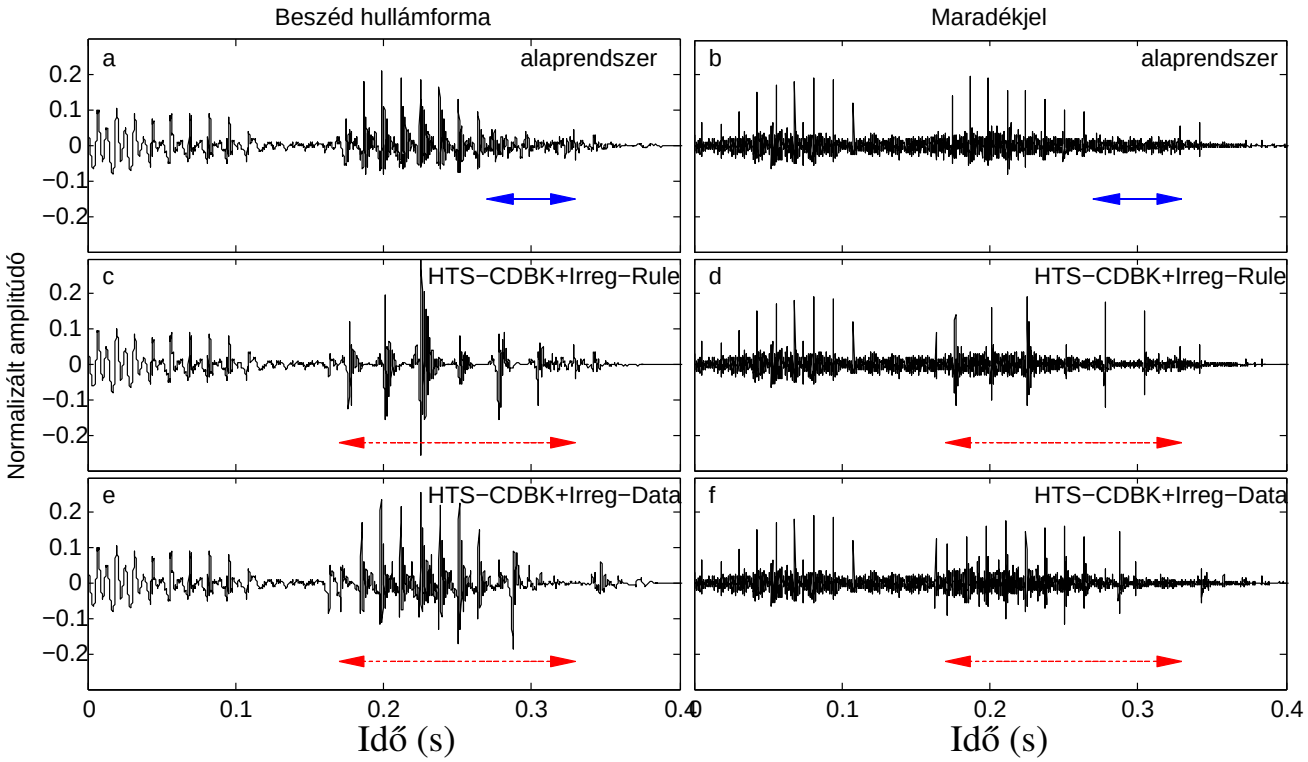
Az analízis és a tanítási lépések a II.1. tézis rendszervével egyezőek, az új rendszer csak a szintézis fázisban különbözik. A kiegészített rendszert HTS-CDBK+Irreg-Rule-nak nevezzük. A HTS-CDBK rendszerhez hasonlóan a glot-

talizáció helyére nincs külön előrejelző eljárás, hanem azt a generált $F0$ paraméterfolyamból állapítjuk meg. Amennyiben legalább 5 egymás utáni magánhangzó keretben nulla az $F0$ értéke, alkalmazzuk az irreguláris zönge modellt az adott magánhangzóra. Ezekben az esetekben az $F0$-menetet lineárisan interpoláljuk a környező zöngés részeknek megfelelően, vagy amennyiben nincs ilyen, akkor enyhén ereszkedő $F0$-menetet állítunk be.

A HTS-CDBK+Irreg-Rule rendszer három heurisztikát használ az irreguláris zönge modellezésére: 1) $F0$ felezés 2) zöngeszinkron maradékjel amplitúdó skálázás véletlen számokkal és 3) spektrális torzítás. A szintézis során a modális zöngés és zöngétlen részeken a HTS-CDBK rendszer által generált maradékjelet használjuk. Azokban a szakaszokban, amelyeknek szintézise irreguláris módon történik, az interpolált $F0$ értékek felét használjuk fel. A glottalizációt gyakran extrém alacsony alapfrekvencia kíséri, a kódkönyvben viszont kevés az ilyen $F0$-lal rendelkező elem. Emiatt a maradékjel periódusokat nullákkal töltjük ki az átlapolt összeadás előtt. Az $F0$ felezés és nullákkal kitöltés eredménye olyan, mintha minden második periódust törölnénk, és ez percepciós szempontból hasonló, mint az alacsonyabb nyitott hányad [5]. A maradékjel szintézisben a kiválasztott kódkönyv elemeken amplitúdó skálázást végzünk: a módszer minden zönge periódust megszoroz egy $\{0 \ldots 1\}$ közötti értékkel. A heurisztika alkalmazását az motiválta, hogy irreguláris zönge esetén az egymás utáni periódusok amplitúdója sokszor ingadozó a kváziperiodikus rezgéssel szemben. Korábbi kutatásban észrevettük, hogy az irreguláris szakaszokon mért MGC paraméterfolyam kevésbé sima a reguláris beszédhez képest (I.2. tézis, [C1]). Emiatt az irreguláris zönge modellben az MGC értékeket torzítjuk: $\{0{,}995 \ldots 1{,}005\}$ közötti véletlen számokkal szorozzuk a paramétereket, ami várhatóan az irreguláris zöngéhez hasonló hatást eredményez. A szintetizált beszédet a maradékjelből a korábbiakhoz hasonlóan MGLSA szűréssel, az MGC paramétereket felhasználva kapjuk vissza.

A 6. ábra egy példát mutat a HTS-CDBK (a és b) és a HTS-CDBK+Irreg-Rule (c és d) rendszerek által generált szóra (vízszintes nyíl jelöli az irreguláris szakaszt). A „Mihály” szó „á” hangjában alkalmaztuk az irreguláris zönge modellt. A nullákkal kitöltés eredményeként a zöngeperiódusok elkülönülnek, míg az amplitúdó skálázás a negyedik periódus erős lecsökkenését eredményezte. Az ábrán látható, hogy a II.2. tézis rendszere jobban hasonlít az eredeti irreguláris beszédre (4. a és 4. b ábra), mint az alaprendszer.

A szabály alapú irreguláris zönge modell eredményének vizsgálatára percepciós tesztet végeztünk. A PPBA adatbázis két férfi beszélőjének (FF3 és FF4) hangja alapján tanítást végeztünk a HTS-CDBK alaprendszerrel és a HTS-CDBK+Irreg-Rule kiegészített rendszerekkel. 130-130 mondatot szintetizáltunk, majd ebből 10-10 olyan mondatot választottunk, amelyben előfordult irreguláris fonáció. A mondatok utolsó, irreguláris magánhangzót tartalmazó szavát kivágtuk és ezeket hasz-



6. ábra. A „Mihály” szó szintetizált változatai (egy hosszabb mondatból kivágva):
a) beszédjel b) maradékjel a II.1. tézis modelljével (alaprendszer)
c) beszédjel d) maradékjel a II.2. tézis modelljével
e) beszédjel f) maradékjel a II.3. tézis modelljével.
Az irreguláris zöngeképzésű szakaszokat vízszintes nyilak jelölik.

náltuk fel a meghallgatásos tesztben. A 11 tesztelő értékelése szerint a II.2. tézis rendszere szignifikánsan kellemesebb hangzású ($p < 0{,}0005$) és szignifikánsan jobban hasonlít az eredeti beszélőre ($p < 0{,}0005$), mint az alaprendszer.

II.3. tézis: [J1] Kidolgoztam egy nyelvfüggetlen adatvezérelt irreguláris zöngeképzés modellt és illesztettem ezt a II.1. tézisben ismertetett gépi szövegfelolvasóhoz. A modell irreguláris beszédrészletek maradékjeléből épített korpuszból elemkiválasztással keresi meg a szintézis során a megfelelő elemeket. Percepciós teszttel igazoltam magyar mintákon, hogy a kiegészített rendszerrel szintetizált beszéd szignifikánsan preferáltabb és jobban emlékeztet az eredeti beszélőre, mint a II.1. tézis rendszere.

Az irreguláris zönge szintézisbe illesztésére egy másik, adatvezérelt modellt is létrehoztunk, amely maradékjel elemkiválasztáson alapul. A kiegészített rendszert HTS-CDBK+Irreg-Data-nak nevezzük. Az analízis és a tanítási lépések a II.1. tézis rendszerével egyezőek, az új rendszer csak a szintézis fázisban különbözik.

Az analízis elvégzése után a beszédadatbázis irreguláris szakaszainak maradékjeléből glottalizációs korpuszt építünk („GLOTT” korpusz). Ehhez egy magas ta-

lálati arányú glottalizáció detektort alkalmazunk („creak_detect”, [20]). Azokat a maradékjel szakaszokat vesszük be a GLOTT korpuszba, amelyek esetén a detektor a magánhangzó kereteinek legalább felében „irreguláris” bináris döntést hozott. Az adatvezérelt módszernél teljes, magánhangzó-hosszúságú maradékjel szakaszokat tárolunk a korpuszban a korábbi zöngeszinkron maradékjel periódusokkal szemben.

A szintézis során a modális maradékjel szakaszok a HTS-CDBK módszerével készülnek. A HTS-CDBK rendszerhez hasonlóan a glottalizáció helyére nincs külön előrejelző eljárás, hanem azt a generált $F0$ paraméterfolyamból állapítjuk meg. Az irreguláris részekhez a glottalizációs korpuszból keresünk illeszkedő elemet. A módszer jelen változatában azt feltételezzük, hogy csak egy magánhangzót kell irreguláris módon szintetizálni, így nem foglalkozunk az elemek közötti összefűzéssel. Az elemkiválasztáshoz csak célköltséget használunk, ami három rész-költségből áll: 1) a paraméterfolyamból származó és a kódkönyv elemek közötti átlagos $F0$ különbség 2) átlagos hossz különbség valamint 3) a maradékjel szakasz hangkörnyezete. Olyan elemeket keresünk, amelyek a szintetizálandó szakasznál hosszabbak. Miután a cél maradékjelet megtaláltuk a célköltség minimalizálásával, a kiválasztott maradékjel utolsó mintáit levágjuk, így beállítva a jeldarab hosszát. Az irreguláris maradékjel energiáját a $gain$ paraméterek átlaga alapján skálázzuk, de a jel más tulajdonságát nem módosítjuk. A HTS-CDBK+Irreg-Rule modellhez hasonlóan spektrális torzítást alkalmazunk, és a végül az MGC paramétereket felhasználó MGLSA szűréssel állítjuk elő a szintetizált beszédet az összefűzött modális és irreguláris maradékjel szakaszokból.

A 6. ábra egy példát mutat az adatvezérelt irreguláris fonáció modell eredményére (e és f). Az alap HTS-CDBK rendszerhez (a és b) hasonlóan a HTS-CDBK+Irreg-Data maradékjele is csak az utolsó magánhangzó egy részében tartalmaz hirtelen amplitúdó ingadozást. Ha ezt összehasonlítjuk az eredeti irreguláris beszédmintával (4. a és 4. b ábra), az látható, hogy a szintetizált maradékjel is másodlagos impulzusokat tartalmaz a periódusokon belül, az eredeti beszéd maradékjelhez hasonlóan.

Percepciós teszttel ellenőriztük az adatvezérelt irreguláris zönge modell eredményét a PPBA adatbázis FF3 és FF4 beszélőinek mintái alapján. Az FF3 beszélő beszédéből 1116 elem, míg az FF4 beszélő adatbázisából 1822 elem került bele a GLOTT korpuszba. 130-130 mondatot szintetizáltunk a HTS-CDBK alaprendszerrel és a HTS-CDBK+Irreg-Data rendszerekkel, majd 10-10 mondatot kiválasztottunk, amelyek tartalmaztak irregulárisan szintetizált magánhangzót, és ezeket a szavakat kivágtuk. A tesztelők a szavak különböző változatait értékelték páros összehasonlítás keretében. A 16 tesztelő eredménye alapján a II.3. tézisben kiegészített rendszer szignifikánsan jobban emlékeztet az eredeti beszélőre

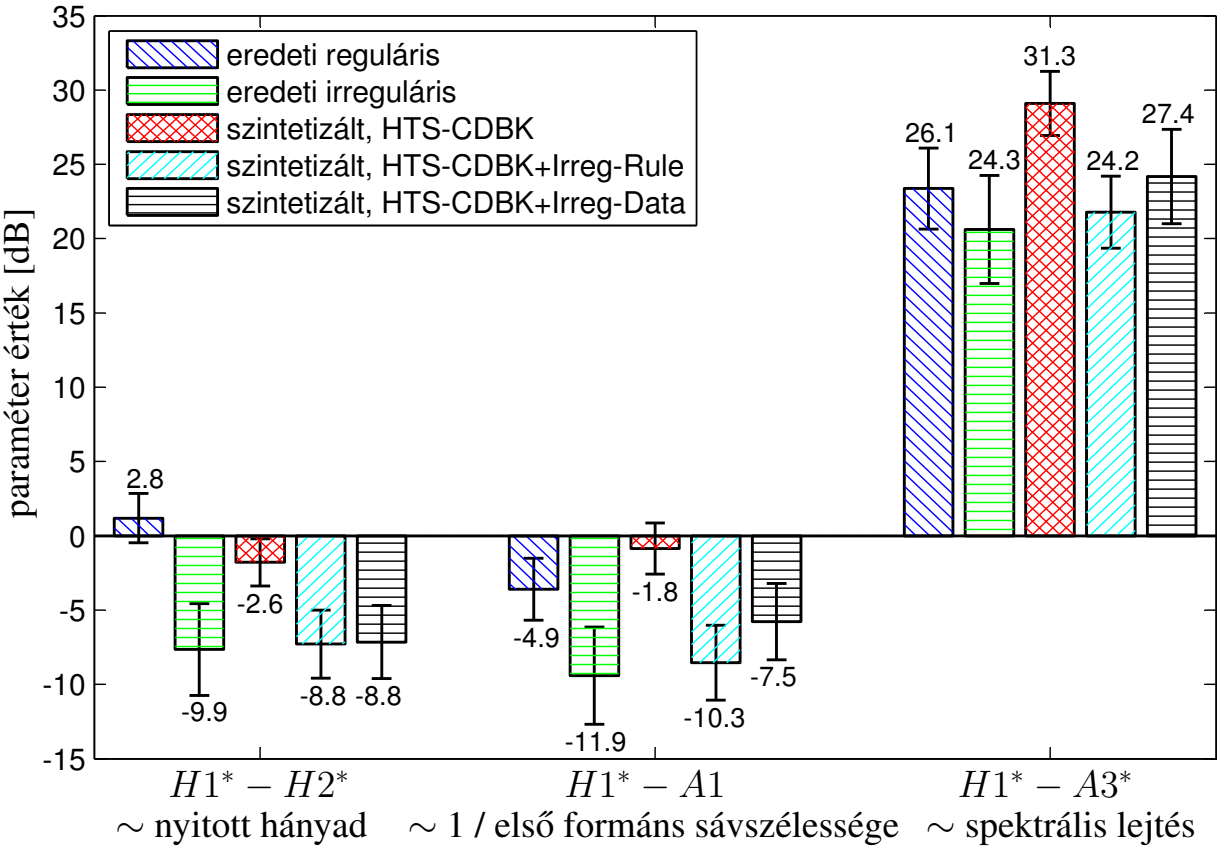
($p < 0{,}0005$) és szignifikánsan jobban preferált ($p < 0{,}0005$), mint az alaprendszer.

II.4. tézis: [J1] Kísérleti úton igazoltam magyar mintákon, hogy a II.2 és II.3. tézisek eljárásai beszédszintézis során a beszéd több releváns akusztikai paraméterét (nyitott hányad: II.2 és II.3, első formáns sávszélessége: II.2) az irreguláris zöngeképzésre jellemző módon modellezik.

A II.2. és II.3. tézisekben meghallgatásos teszthez kiválasztott beszédmintákon akusztikus elemzést is végeztünk, az I.3. tézishez hasonló módon. A szakirodalom alapján kiválasztottunk három olyan akusztikai jegyet, amelyeket korábban irreguláris és reguláris beszéd megkülönböztetésére használtak [30, 5]. Ezek alapján irreguláris beszédben a nyitott hányad (OQ) alacsonyabb; az első formáns sávszélessége ($B1$) nagyobb; a spektrum lejtése (TL) meredekebb, mint reguláris beszédben.

Az elemzéseket a két beszélő 10-10 szintetizált szaván, és 10-10 másik, eredeti reguláris és eredeti irreguláris felvételen végeztük. Az OQ helyett a $H1^* - H2^*$-ot mértük, az $1/B1$-et $H1^* - A1$ elemzésével vizsgáltuk, a TL akusztikai jegyet pedig $H1^* - A3^*$ alapján mértük. A három paraméter és öt beszédminta típus összehasonlítása a 7. ábrán látható. A Tukey-HSD post-hoc teszttel kiegészített ANOVA elemzés szerint az első két harmonikus különbsége szignifikánsan különbözik az eredeti reguláris beszéd, szintetizált alaprendszer és a többi beszédminta között ($p < 0{,}05$), azonban nem tér el jelentősen az eredeti és szintetizált irreguláris változatok között ($p = 0{,}99$, n.s.). Eszerint mindkét irreguláris fonáció modell megfelelően modellezi a nyitott hányadot. A 7. ábra alapján a $H1^* - A1$ és így az első formáns sávszélességének szempontjából a szabály alapú irreguláris zönge modell közel áll az eredeti irreguláris beszédhez, míg az adatvezérelt modell eredménye a reguláris és irreguláris minták között helyezkedik el. Ebben a kísérletben a $H1^* - A3^*$ érték nem segítette a beszédminták elkülönítését. Az akusztikus elemzésből azt a következtetést vonhatjuk le, hogy a vizsgált három jellemző közül kettő esetén a szintetizált változatok közel vannak az eredeti glottalizált beszédhez.

Összességében a percepciós tesztek és az akusztikus elemzés alapján azt mondhatjuk, hogy a mindkét irreguláris zönge modell alkalmas glottalizált beszéd szintézisére, és a rendszerekkel létrehozott beszédminták minősége hasonló.



7. ábra. Az irreguláris zönge modellekkel szintetizált szavak akusztikus elemzésének eredménye. A függőleges fekete vonalak a 95%-os konfidenciaintervallumot jelölik.

III. téziscsoport: Szubglottális rezonanciák elemzése a magyar beszédben

A beszédkeltés forrás-szűrő modellje [1], melyet az I.1. tézis gerjesztési modelljének kidolgozása során is alkalmaztunk, azt az egyszerűsítést használja, hogy a forrás és a szűrő tökéletesen szétválasztható. A valóságban azonban a forrás (gége) és a szűrő (artikulációs csatorna) között nemlineáris csatolás jöhet létre, melyet részben a szubglottális rendszer okoz. A formánsok (az artikulációs csatorna rezonancia frekvenciái) és a szubglottális rezonanciák (az alsó légúti tér rezonancia frekvenciái) között ugyan nincs közvetlen ok-okozati összefüggés, azonban a közöttük fennálló indirekt kapcsolat különböző magánhangzó csoportok elkülönüléséhez vezet, melyre a kvantális elmélet ad magyarázatot. A kvantális elmélet [25] elvileg univerzálisan, nyelvektől függetlenül rendszert alkot a beszédhangok kategorizálására, azonban a gyakorlatban nem egyértelmű, hogy a szubglottális rezonanciák minden nyelven hozzájárulnak-e a beszédhangok elkülönítéséhez. A szubglottális rezonanciák és magánhangzó formánsok kapcsolatát korábban vizsgálták beszédprodukciós szempontból amerikai angol [7], spanyol, német és koreai nyelvre; magyarra azonban eddig voltak eredmények.

Ebben a téziscsoportban bemutatom a szubglottális rezonanciák vizsgálatára irányuló magyar nyelvre végzett elemzéseimet és egy új, szubglottális rezonancia alapú magánhangzó osztályozó eljárást.

III.1. tézis: [C4, J4] Modellt dolgoztam ki az alsó légúti (szubglottális) rendszer rezonanciáinak magyar beszédre vonatkozó hatására. Kimutattam, hogy a szubglottális rezonanciák (az alsó légúti rendszer első három rezonanciafrekvenciája) magyar beszédben felhasználhatóak magánhangzó osztályok formánsok szerinti elkülönítéséhez a szubglottális rezonanciák és formánsok közti indirekt kapcsolatot kihasználva.

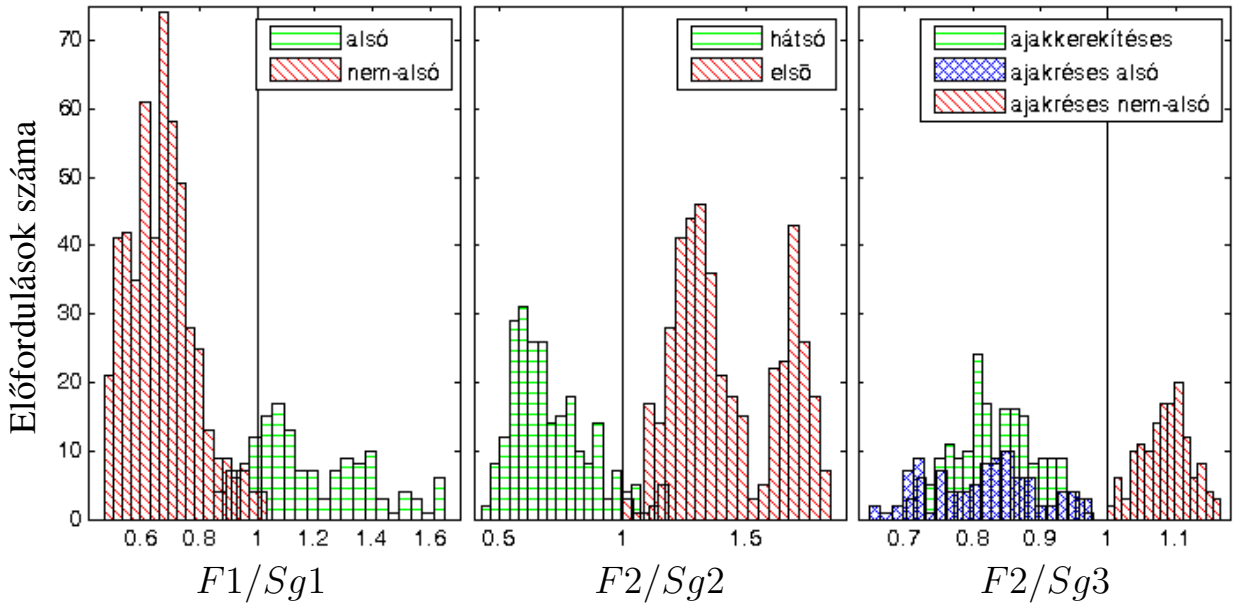
Első kísérletként magyar magánhangzókra vizsgáltuk a szubglottális rendszer hatását. Ehhez új felvételeket rögzítettünk, melyek alapján beszélőnként különkülön és összevonva, normalizálással is végeztünk elemzéseket.

A kutatás során négy magyar anyanyelvű beszélő beszéd és gyorsulásmérő jelét elemeztük logatom-felolvasásban. Az első három formáns értékét ($F1$, $F2$ és $F3$) automatikusan mértük a beszédjelből a Praat programmal a vizsgálandó magánhangzók közepén, majd manuálisan javítottuk. A szubglottális rezonanciákat ($Sg1$, $Sg2$ és $Sg3$) manuálisan mértük a Wavesurfer programmal a gyorsulásmérő jelből az LPC spektrum burkolójának csúcsaiként, minden beszélő és SGR esetén 25-25 ponton. A beszélők szubglottális rezonanciáinak mediánjait használtuk fel a továbbiakban. A mérések alapján a szubglottális rezonanciák beszédre vonatkozó hatására akusztikai alapú modellt dolgoztunk ki. Az amerikai angol nyelvre kidolgozott modellt [7] alkalmaztuk a magyar nyelvre, és megállapítottuk, hogy

1) az első szubglottális rezonancia ($Sg1$) az első formáns ($F1$) tartományában az alsó és a nem-alsó nyelvállású magánhangzók között található,
2) a második szubglottális rezonancia ($Sg2$) a második formáns ($F2$) tartományában az elöl és hátul képzett magánhangzók között található,
3) a harmadik szubglottális rezonancia ($Sg3$) a második formáns ($F2$) tartományában az elöl képzett, ajakréses, nem-alsó magánhangzókat választja el a többi elöl képzett magánhangzótól.

A fenti modellt mérnöki módszerekkel igazoltuk: a magánhangzó formánsok normalizálásával az egyes formáns értékeket a beszélő megfelelő szubglottális rezonanciájával elosztottuk ($F1/Sg1$, $F2/Sg2$ és $F2/Sg3$), majd a beszélőnkénti adatokat összevontuk. A 8. ábra az SGR-normalizált formáns hisztogramokat mutatja: például a b) ábrán az vehető észre, hogy az $Sg2$ (függőleges vonal) az elöl és a hátul képzett magánhangzókat közel optimálisan választja el.

A részletes vizsgálatok szerint a fentiek nem teljesülnek minden beszélő és minden kategória esetén. A kategóriák optimális elválasztásának részletes vizsgálatára



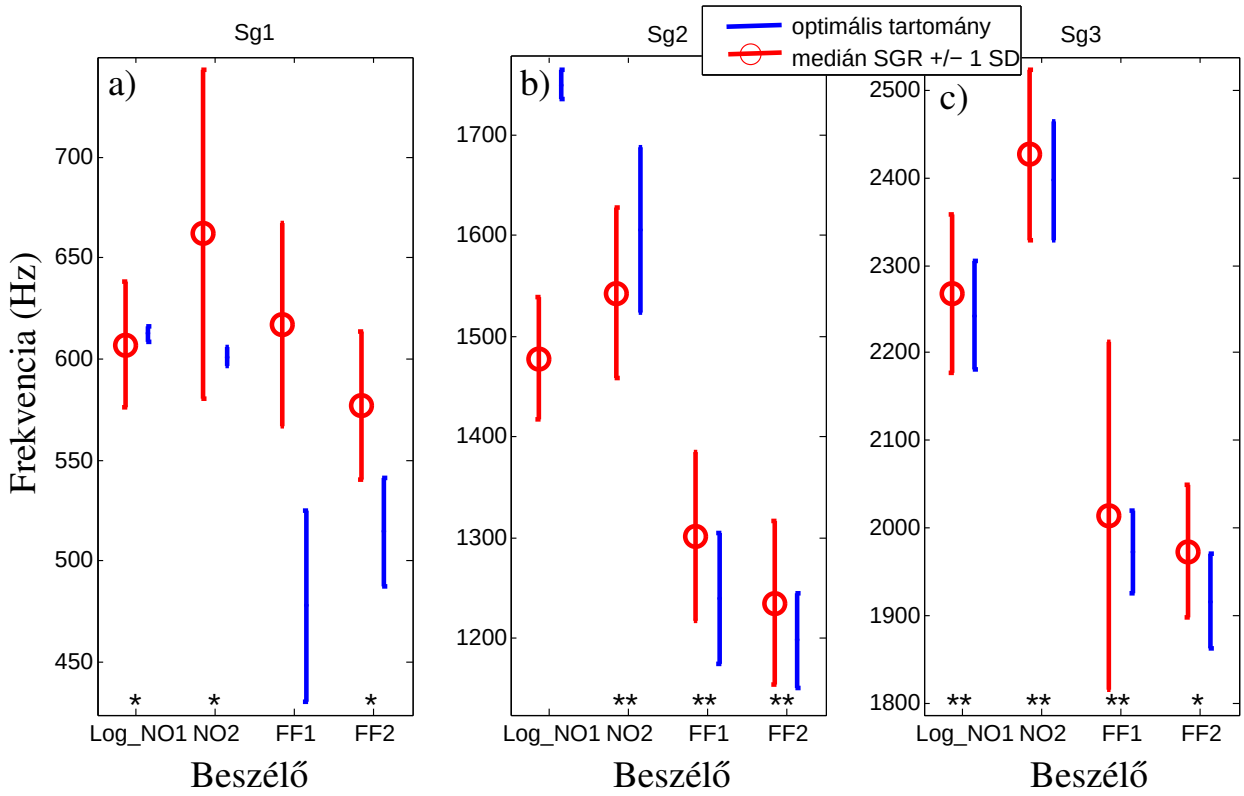
8. ábra. Normalizált formáns hisztogramok logatom beszéd alapján: az $F1/Sg1$, $F2/Sg2$, $F2/Sg3$ értékek összevonva az összes beszélőre. A függőleges vonal a normalizált $Sg1$, $Sg2$, $Sg3$ értéket jelöli.

ROC (*Receiver Operating Characteristics*) elemzést végeztünk külön-külön minden SGR-re és beszélőre, amelynek eredménye a 9. ábrán látható. Az elemzés megmutatta, hogy a 12-ből 6 esetben az SGR mediánja az optimális elválasztási tartományon belül van (** az ábrán), 4 további esetben egységnyi szóráson belül található (*), míg a maradék 2 esetben távolabb van.

Összefoglalva az eredményeket, a szubglottális rezonanciák közel optimálisan választják el egymástól az alsó vs. nem-alsó nyelvállású, elöl képzett vs. hátul képzett, illetve elöl képzett, ajakréses, nem-alsó nyelvállású vs. egyéb elöl képzett magánhangzókat a magyar nyelvben.

A III.1. tézisben bemutatott formális, kvantitatív modell alkalmas gépi implementációra. Annak tesztelésére, hogy a szubglottális rezonanciák ismerete segítheti-e a beszédfeldolgozást, egy kísérletet terveztünk, amelyben megvizsgáljuk, hogy az SGR-ek felhasználásával pontosabb osztályozást tudunk-e végezni magánhangzókra, mint anélkül.

III.2. tézis: [J4] Automatikus osztályozót készítettem, mely egy beszélő magánhangzó formánsainak és szubglottális rezonanciáinak indirekt kapcsolatán alapulva normalizálásával a magánhangzókat a III.1. tézisben ismertetett kategóriákba sorolja. Megmutattam, hogy a vizsgált mintákon az Sg2 alapú módszer mindig pontosabb, az Sg3 alapú módszer kis tanítóadat-mennyiség esetén pontosabb, míg az Sg1 alapú módszer nem pontosabb mint egy tisztán formánsokat felhasználó döntési fa alapú referencia osztályozó.



9. ábra. ROC elemzés eredménye a szubglottális rezonanciák magánhangzó csoportokra elkülönítésének vizsgálatára. A világos vonalak az SGR értékeket és egységnyi szórásukat mutatják, a sötét vonalak az optimális elválasztó tartományt jelölik.

Tanító és tesztelő adatként hat magyar anyanyelvű beszélő (5 férfi és 1 nő) spontán beszéd felvételeiből [28] származó 5948 magánhangzót használtunk fel. Az $Sg1$, $Sg2$ és $Sg3$ értékeket külön olvasott beszéd felvételek alapján kézzel mértük a Wavesurfer programban minden beszélő és SGR esetén 20-20 ponton, majd a mediánjaikat használtuk fel.

A kísérlet során referencia osztályozónak J4.8 típusú döntési fákat használtunk a Weka programban. A döntési fára azért esett a választás, mert ez a C4.5 típusú, széles körben használt döntési fa továbbfejlesztett változata, és a legtöbb esetben közel optimális osztályozást eredményez. A három szubglottális rezonanciának és a III.1. tézisben ismertetett modellnek megfelelően három osztályozót készítettünk, melyek bemenetei a magánhangzónkénti tiszta formáns értékek, kimenetei pedig a modell kategóriái:

a) bemenet: $F1$, kimenet: alsó – nem-alsó
b) bemenet: $F2$, kimenet: elöl képzett – hátul képzett
c) bemenet: $F2$, kimenet: elöl képzett, ajakréses, nem-alsó – egyéb

Ezután a három szubglottális rezonanciának és a III.1. tézisben ismertetett modell három kategóriájának megfelelően SGR alapú formáns normalizálást használó osztályozókat készítettünk. Az osztályozók bemenete a magánhangzó $F1$ vagy

$F2$ formánsának normalizált értéke, azaz a formánsfrekvencia elosztva a megfelelő szubglottális rezonancia frekvenciájával ($Sg1$, $Sg2$ vagy $Sg3$). Az osztályozók kimenetei a modellben ismertetett magánhangzó kategóriák:

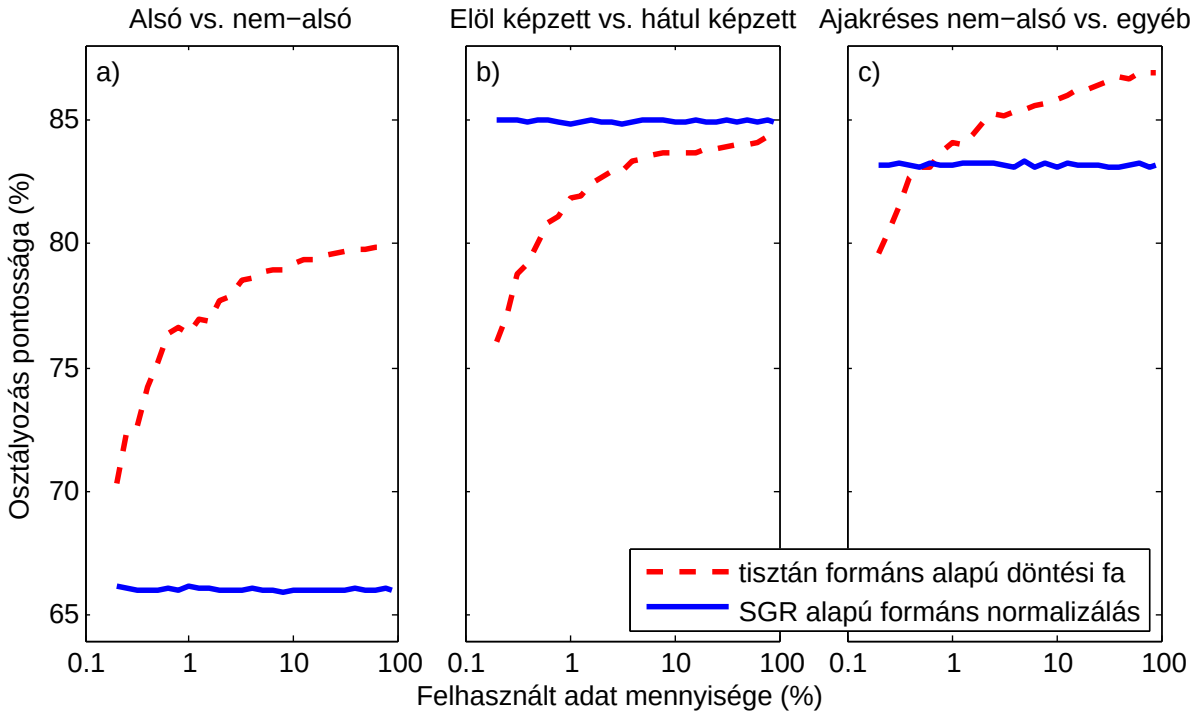
a) bemenet: $Fn1 = F1/Sg1$, kimenet: alsó – nem-alsó
b) bemenet: $Fn2 = F2/Sg2$, kimenet: elöl képzett – hátul képzett
c) bemenet: $Fn3 = F2/Sg3$, kimenet: elöl képzett, ajakréses, nem-alsó – egyéb

Az osztályozó egyszerű küszöbérték alapján működik: például a b) esetben amennyiben a bemeneti magánhangzóra vonatkozó $Fn2 \geq 1{,}0$, akkor elöl képzett kategóriára dönt, ha $Fn2 < 1{,}0$, akkor pedig hátul képzett kategóriára dönt az osztályozó.

A tiszta formáns bemenetet használó (SGR-ek ismerete nélküli) referencia osztályozókat összehasonlítottuk a szubglottális rezonancia-normalizálás alapú osztályozók eredményével. A kísérletben azt vizsgáltuk, hogy a felhasznált adat mennyiségének függvényében melyik osztályozó teljesít jobban. A döntési fa alapú osztályozó esetén a tanítóadatot a teljes adat 0,2...90 %-a között (12 – 5353 adatpont) változtattuk, és a maradék adatmennyiséget használtuk tesztelésre. Az SGR alapú osztályozó pontossága nem függ a tanító adat mennyiségétől, amennyiben a szubglottális rezonancia értékek meghatározásra kerültek. A szubglottális rezonancia alapú osztályozás esetén az adathalmaz 50 %-án végeztük a teszteket. Minden mérést 100 véletlen csoporton ismételtünk és az eredményeket átlagoltuk.

A kísérlet eredményeit a 10. ábra mutatja. Az a) esetben az $Sg1$ ismerete nem segítette az osztályozást. Ezt valószínűleg az okozta, hogy az $Sg1$ mérése sokszor nehézkes a gyorsulásmérő felvételből, mert az intenzív alsó harmonikusok torzíthatják a méréseket. A b) ábrán az elöl képzett és hátul képzett magánhangzók elkülönítésének eredménye látható. Ebben az esetben az $Sg2$ ismerete egyértelműen javította az osztályozást: kevés adat esetén közel 20 %-kal, míg az adathalmaz jelentős részét felhasználva is 1 %-kal pontosabb a szubglottális rezonancia alapú osztályozó a tisztán formáns alapú döntési fához képest. Ez az eredmény megfelel a szakirodalom alapján elvártnak, mert a kutatások szerint a szubglottális rezonanciák közül az $Sg2$ hatása a legjelentősebb a magánhangzók kategóriákra osztásában [7]. A c) esetben az $Sg3$ alapú osztályozás pontosabb, amennyiben átlagosan az adathalmaz kevesebb, mint 1 %-át (50 magánhangzó) ismerjük.

A kísérlet alapján az $Sg2$ alapú osztályozás mindig, míg az $Sg3$ kevés tanító adat rendelkezésre állása esetén (50 magánhangzónál kevesebb adat) jobb eredményre vezet a döntési fa alapú referencia osztályozónál. Az SGR-normalizálás alapú osztályozás esetén elegendő körülbelül 10-20 magánhangzó tanító adatnak, melyek a szubglottális rezonanciák méréséhez szükségesek. Az SGR alapú módszer tehát gyorsan adaptálódik a beszélőhöz, és elméletileg megalapozott, mivel



10. ábra. A tisztán formáns alapú döntési fa, és SGR-normalizált formáns alapú automatikus osztályozók pontosságának összehasonlítása a tanításhoz felhasznált adat mennyiségének függvényében: a) $Sg1$, b) $Sg2$, c) $Sg3$.

a III.1. tézis modellje alapján működik. A tisztán formáns alapú módszer viszont érzékeny a tanítóminták jellegére és mennyiségére.

A fentiek során a beszédhangok formánsainak és a szubglottális rezonanciáknak az összefüggését vizsgáltuk beszédprodukcióban és automatikus osztályozás során, magyar nyelvre. Az elemzések és kísérletek szerint a szubglottális rezonanciák magyar nyelven is segítik a magánhangzók fonológiai megkülönböztető jegyek szerinti elkülönülését, így hozzájárulva a kvantális elmélet [25] szubglottális rezonanciákra vonatkozó kiegészítéséhez [7].

6. Az eredmények alkalmazhatósága

Kutatásom eredményei számos beszédtechnológiai alkalmazásban felhasználhatóak, amelyek egyrészt hozzájárulhatnak a természetesebb ember-gép kommunikációhoz, másrészt segíthetnek megérteni az emberi beszédképzés működését. Eljárásaimat magyar nyelvű mintákon teszteltem. Az I. és II. téziscsoportok módszerei nyelvfüggetlenek, így a modellek kiterjeszthetőek más nyelvekre is. Az alábbiakban téziscsoportonként bemutatok néhány alkalmazási lehetőséget.

Az I.1. tézisben ismertetett maradékjelen alapuló analízis-szintézis gerjesztési modell alkalmas különböző zöngeminőségek gépi előállítására és transzformációjára. Előzetes kísérleteim szerint levegősből modális beszéd átalakítására is meg-

felelő lehet a módszer. Az I.2. tézis glottalizáció javító eljárását ki lehet terjeszteni hosszabb beszédszakaszokra is, amivel rekedtes, patologikus hangokat várhatóan szebbé, kellemesebbé lehet tenni (pl. színészek, bemondók hangja). Az irreguláris-reguláris átalakító eljárás automatikussá kiegészített változatával beszédadatbázisokból el lehetne tüntetni az irreguláris zöngéjű szakaszokat, ezáltal ideálisabbá téve a beszédet a további feldolgozás céljából.

A II.1. tézisben bemutatott beszédszintetizátor rendszer javíthatja a korlátozott erőforrású eszközökben (pl. okostelefon) alkalmazott gépi szövegfelolvasás minőségét. A kevés erőforrás miatt bonyolultabb gerjesztési modellek nehézkesen kezelhetőek, viszont a tézis modellje várhatóan bizonyos korlátozott erőforrású eszközökön képes valós idejű működésre. A II.2. és II.3. tézisek irreguláris zönge modelljei hozzájárulhatnak a természetesebb, expresszív és személyre szabott beszédszintézishez. A természetességen és személyre szabhatóságon itt azt értem, hogy az eredeti beszédadatbázisban előforduló glottalizált eseteknek megfelelő arányú irreguláris hangot tudunk képezni szintetizált beszédben is. Korábban kimutatták, hogy bizonyos érzelmeket (pl. szomorú és ingerült) a beszélők a zönge minőség módosításával is jeleznek; így az irreguláris zönge modell javíthatja az érzelmes, expresszív beszédszintézist.

A szubglottális rezonanciák vizsgálata, így a III.1. tézis hozzájárul a kvantális elmélet szerinti fonológiai megkülönböztető jegyek működésének megértéséhez. A feltételezések szerint a percepció során a beszédhangok formánsait részben a szubglottális rezonanciákhoz viszonyítjuk (normalizáljuk), ezáltal megkönnyítve egymás beszédének megértését, hiszen az egyes egyének akusztikai produktumában nagy eltérések mutatkoznak. Ez a tulajdonság kihasználható a beszédtechnológiában is: a szubglottális rezonanciákat már sikerrel alkalmazták beszédfelismerő rendszer javítására gyermek beszéd esetén [26]. Egy előzetes percepciós teszt során megfigyeltük, hogy a formánsok és szubglottális rezonanciák aránya kapcsolatba hozható az észlelt magánhangzó minőségével [J4]. A kísérlet eredményei szerint várhatóan a nem megfelelő $F2 - Sg2$ arány (azaz amennyiben a kapcsolat nem a III.1. tézis modellje szerinti) a természetes beszéd során percepciós szempontból előnytelen, és nehezíti a beszéd megértését. Ez alapján készíthető egy olyan eljárás, amely beszédszintetizátor adatbázisából kitisztítja a formáns – szubglottális rezonancia szempontjából nem megfelelő beszédrészleteket, ezzel hozzájárulva a szintetizált beszéd érthetőbbé tételéhez. A III.2. tézisben ismertetett osztályozó kiegészíthető hosszabb hangkapcsolatok (pl. CV vagy VC kapcsolat) artikuláció szerinti osztályozására is, melyre amerikai angol nyelvű mintákkal készült már kísérlet. Amennyiben a rejtett Markov-modell alapú beszédszintetizátorban a forrás-szűrő modellt sikerül kiegészíteni a szubglottális rezonanciák modellezésével, az tovább javíthatja a gépi beszéd természetességét.

7. Köszönetnyilvánítás

Ezúton mondok köszönetet konzulensemnek, Dr. Németh Gézának témavezetéséért, a munkám során nyújtott folyamatos segítségéért és támogatásáért, hasznos tanácsaiért és észrevételeiért. Köszönöm neki, hogy munkájával megalapozta tudományos szemléletemet.

Köszönettel tartozom a Beszédtechnológiai Laboratórium jelenlegi és volt munkatársainak. Bartalis Mátyás baráti beszélgetésekkel, Dr. Bőhm Tamás kutatási és módszertani irányelvekkel, Dr. Fék Márk beszédkódolással kapcsolatos ismereteivel, Kiss Géza programozási segítséggel, Dr. Olaszy Gábor nagymértékű tapasztalatával, Tóth Bálint a statisztikai parametrikus beszédszintézis megismertetésével, Dr. Zainkó Csaba jelfeldolgozási ismereteivel segítette munkámat és járult hozzá a disszertáció létrejöttéhez. Emellett köszönöm Fegyó Tibor, Kiss Gábor, Dr. Mihajlik Péter, Nagy Péter, Dr. Szaszák György, Sztahó Dávid, Tarján Balázs és Dr. Vicsi Klára segítségét.

Köszönöm Dr. Steven M. Lulichnak (Indiana University, Bloomington, USA), hogy megismertette velem szubglottális rezonanciákkal foglalkozó kutatásait és támogatta kísérleteimet ebben a témában. Köszönettel tartozom Dr. Gráczi Tekla Etelkának (MTA Nyelvtudományi Intézet), Dr. Bárkányi Zsuzsannának (MTA Nyelvtudományi Intézet) és Beke Andrásnak (MTA Nyelvtudományi Intézet) a kutatási együttműködésért és látóköröm szélesítéséért. Köszönöm minden társszerzőmnek a közös cikkek írásának lehetőségét és a csapatmunkában történő kutatás örömét.

Köszönöm továbbá Dr. Henk Tamás és Dr. Magyar Gábor tanszékvezető uraknak, hogy vezetésük alatt a tanszéken végezhettem doktori munkámat.

A PPBA, BEA adatbázisok és a III. téziscsoport beszélőinek köszönöm, hogy a kísérleteimhez felhasználhattam a hangjukat. A percepciós tesztekben résztvevőknek köszönöm, hogy meghallgatták és értékelték a hanganyagokat, valamint megjegyzéseikkel a kutatási irányok távlati meghatározásában is segítettek.

Köszönöm Dr. Gósy Máriának és Dr. Olaszi Péternek, hogy értékes észrevételeikkel és hasznos javaslataikkal segítették a disszertáció jobbá tételét.

Külön köszönöm családomnak: feleségemnek Berninek, kislányomnak Lilinek, kisfiamnak Ábelnek, édesanyámnak Édinek, édesapámnak Istvánnak és bátyámnak Krisztiánnak, hogy doktori tanulmányaim alatt folyamatosan támogattak és megteremtették számomra a kutatáshoz szükséges nyugodt légkört.

A kutatást a NAP (OMFB-00736/2005), az Enhances (NKFP 2/034/2004), a Teleauto (OM-00102/2007), a BelAmi (ALAP2-00004/2005), az ETOCOM (TÁMOP-4.2.2-08/1/KMR-2008-0007), a Kutatóegyetem (TÁMOP-4.2.1/B-09/1/KMR-2010-0002), a CESAR (Grant No. 271022), a Paelife (Grant No. AAL-08-1-2011-0001) és az EITKIC_12-1-2012-001 projektek támogatták.

8. Rövidítések

ANOVA	ANalysis Of VAriance / Varianciaanalízis
BEA	BEszélt nyelvi Adatbázis
C	Consonant / Mássalhangzó
CELP	Code-Excited Linear Prediction
CMOS	Comparative Mean Opinion Score
DSM	Deterministic plus Stochastic Model / Determinisztikus-sztochasztikus modell
GCI	Glottal Closure Instant
HMM	Hidden Markov-model / Rejtett Markov-modell
HNM	Harmonic plus Noise Model / Harmonikus-zaj modell
HNR	Harmonics-To-Noise Ratio / Harmonikus-zaj arány
HTS	HMM-based Speech Synthesis System (H-Triple-S)
IPA	International Phonetic Alphabet / Nemzetközi Fonetikai Ábécé
LF	Liljencrants-Fant
LPC	Linear Predictive Coding / Lineáris Predikciós Kódolás
MGC	Mel-Generalized Ceptstrum / Mel-Általánosított Kepsztrum
MGLSA	Mel-Generalized Log Spectral Approximation
MOS	Mean Opinion Score
MSD	Multi-Space Distribution / Többterű eloszlás
OQ	Open Quotient / Nyitott hányad
PCA	Principal Component Analysis / Főkomponensanalízis
PN	Pulse-Noise / Impulzus-zaj
PPBA	Preciziós, Párhuzamos magyar Beszédadatbázis
PSOLA	Pitch Synchronous Overlap and Add / Zöngeszinkron átlapoló összegzés
QT	Quantal Theory / Kvantális elmélet
RMS	Root Mean Square / Négyzetes átlag
RMSE	Root Mean Squared Error / Átlagos négyzetes hiba
ROC	Receiver Operating Characteristics
SEDREAMS	Speech Event Detection using the Residual Excitation And a Mean-based Signal
SGR	Subglottal Resonance / Szubglottális rezonancia
TL	Spectral Tilt / Spektrális lejtés
TTS	Text-To-Speech / Gépi szövegfelolvasás
V	Vowel / Magánhangzó

9. Jelölések

A1	Első formáns amplitúdója
A3, A3*	Harmadik formáns amplitúdója (*: korrigált érték)
B1	Első formáns sávszélessége
F0	Alapfrekvencia
F1	Első formáns frekvenciája
F2	Második formáns frekvenciája
F3	Harmadik formáns frekvenciája
FF1, FF2, FF3, FF4	PPBA adatbázis négy férfi beszélője
Fn1	Sg1-normalizált első formáns
Fn2	Sg2-normalizált második formáns
Fn3	Sg3-normalizált második formáns
H1, H1*	Első harmonikus (*: korrigált érték)
H2, H2*	Második harmonikus (*: korrigált érték)
HTS-CDBK	Maradékjel kódkönyv gerjesztésű HTS
HTS-CDBK+Irreg-Rule	Szabály alapú irreguláris zönge modell HTS-ben
HTS-CDBK+Irreg-Data	Adatvezérelt irreguláris zönge modell HTS-ben
HTS-HUN	A HTS rendszer magyar nyelvű változata
HTS-PN	Impulzus-zaj gerjesztésű HTS
gain	Maradékjel periódus energiája
Log_FF1, Log_FF2	Logatom felvételek két férfi beszélője
Log_NO1, Log_NO2	Logatom felvételek két női beszélője
NO3	PPBA adatbázis egyik női beszélője
rt0	Maradékjel periódus csúcsok leírásának paramétere
Sg1	Első szubglottális rezonancia
Sg2	Második szubglottális rezonancia
Sg3	Harmadik szubglottális rezonancia
Spo_FF1 ... Spo_FF5	Spontán beszéd felvételek öt férfi beszélője
Spo_NO1	Spontán beszéd felvételek egy női beszélője

10. Hivatkozások

[1] G. Fant, *Acoustic theory of speech production*. The Hague: Mouton, 1960.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, and A. Black, „The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, (Bonn, Germany), pp. 294–299, 2007.

[3] H. Zen, K. Tokuda, and A. W. Black, „Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, Nov. 2009.

[4] A. Hunt and A. Black, „Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, vol. 1, (Atlanta, Georgia, USA), pp. 373–376, 1996.

[5] T. Bőhm, N. Audibert, S. Shattuck-Hufnagel, G. Németh, and V. Aubergé, „Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles," in *Acoustics'08*, (Paris, France), pp. 6141–6146, 2008.

[6] K. N. Stevens, *Acoustic Phonetics*. Cambridge: Cambridge University Press, 1998.

[7] S. M. Lulich, „Subglottal resonances and distinctive features," *Journal of Phonetics*, vol. 38, no. 1, pp. 20–32, 2010.

[8] B. Tóth and G. Németh, „Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis," *Acta Cybernetica*, vol. 19, no. 4, pp. 715–731, 2010.

[9] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, „Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 1, pp. 325–333, 2007.

[10] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, „Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech*, (Brisbane, Australia), pp. 1829–1832, 2008.

[11] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, „HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. Interspeech*, (Brisbane, Australia), pp. 1881–1884, 2008.

[12] T. Raitio, A. Suni, M. Vainio, and P. Alku, „Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. ICASSP*, (Vancouver, Canada), pp. 7830–7834, 2013.

[13] D. Erro and I. n. Sainz, „HNM-based MFCC+ F0 extractor applied to statistical speech synthesis," in *Proc. ICASSP*, (Prague, Czech Republic), pp. 4728–4731, 2011.

[14] Z. Wen and J. Tao, „Inverse Filtering Based Harmonic plus Noise Excitation Model for HMM-based Speech Synthesis," in *Proc. Interspeech*, (Florence, Italy), pp. 1805–1808, 2011.

[15] J. S. Sung, D. H. Hong, H. W. Koo, and N. S. Kim, „Statistical Approaches to Excitation Modeling in HMM-Based Speech Synthesis," *IEICE Transactions on Information and Systems*, vol. E96-D, no. 2, pp. 379–382, 2013.

[16] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, „Using a Pitch-Synchronous Residual Codebook for Hybrid HMM/frame Selection Speech Synthesis," in *Proc. ICASSP*, (Taipei, Taiwan), pp. 3793 – 3796, 2009.

[17] T. Drugman, G. Wilfart, and T. Dutoit, „A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, (Brighton, UK), pp. 1779–1782, 2009.

[18] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, „Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *The Journal of the Acoustical Society of America*, vol. 103, pp. 2649–2658, May 1998.

[19] T. Bőhm, Z. Both, and G. Németh, „Automatic Classification of Regular vs. Irregular Phonation Types," in *NOLISP*, (Vic, Spain), pp. 43–50, 2009.

[20] J. Kane, T. Drugman, and C. Gobl, „Improved automatic detection of creak," *Computer Speech & Language*, vol. 27, pp. 1028–1047, June 2013.

[21] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, „Parameterization of vocal fry in HMM-based speech synthesis," in *Proc. Interspeech*, (Brighton, UK), pp. 1775–1778, 2009.

[22] T. Drugman, J. Kane, and C. Gobl, „Modeling the Creaky Excitation for Parametric Speech Synthesis," in *Proc. Interspeech*, (Portland, Oregon, USA), pp. 1424–1427, 2012.

[23] T. Drugman, J. Kane, T. Raitio, and C. Gobl, „Prediction of Creaky Voice from Contextual Factors," in *Proc. ICASSP*, (Vancouver, Canada), pp. 7967–7971, 2013.

[24] T. Raitio, J. Kane, T. Drugman, and C. Gobl, „HMM-based synthesis of creaky voice," in *Proc. Interspeech*, pp. 2316–2320, 2013.

[25] K. N. Stevens, „On the quantal nature of speech," *Journal of Phonetics*, vol. 17, pp. 3–45, 1989.

[26] S. Wang, S. M. Lulich, and A. Alwan, „Automatic detection of the second subglottal resonance and its application to speaker normalization," *The Journal of the Acoustical Society of America*, vol. 126, pp. 3268–3277, Dec. 2009.

[27] G. Olaszy, „Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai," *Beszédkutatás 2013*, pp. 261–270, 2013.

[28] M. Gósy, „Magyar spontánbeszéd-adatbázis - BEA," *Beszédkutatás 2008*, pp. 194–207, 2008.

[29] G. de Krom, „A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech and Hearing Research*, vol. 36, pp. 254–266, Apr. 1993.

[30] D. H. Klatt and L. C. Klatt, „Analysis, synthesis, and perception of voice quality variations among female and male talkers.," *The Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, Feb. 1990.

[31] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. C. Guiod, and S. L. Goldman, „Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *Journal of Speech and Hearing Research*, vol. 38, pp. 1212–1223, Dec. 1995.

[32] M. Iseli and A. Alwan, „An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proc. ICASSP*, (Montreal, Quebec, Canada), pp. 669–672, 2004.

11. Publikációs tevékenység

A tézispontokhoz kapcsolódó tudományos közlemények

Folyóiratcikkek

[J1] Tamás Gábor Csapó, Géza Németh, „Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation," *IEEE Journal on Selected Topics in Signal Processing*, elfogadva, 2013.
(BME-PA pontszám: 100% · 6p = 6p.) Scopus / Web of Science, IF: 3.297.

[J2] Tamás Gábor Csapó, Géza Németh, „Statistical parametric speech synthesis with a novel codebook-based excitation model," *Intelligent Decision Technologies*, elfogadva, 2013.
(BME-PA pontszám: 100% · 6p = 6p.) Scopus.

[J3] Tamás Gábor Csapó, „Increasing the naturalness of synthesized speech (PhD summary)," *The Phonetician*, No. 104–105, pp. 88–97, 2012.
(BME-PA pontszám: 100% · 0p = 0p.) (ismeretterjesztő cikk)

[J4] Tamás Gábor Csapó, Tekla Etelka Gráczi, Zsuzsanna Bárkányi, András Beke, Steven M. Lulich, „Patterns of Hungarian vowel production and perception with regard to subglottal resonances," *The Phonetician*, No. 99–100, pp. 7–28, 2011.
(BME-PA pontszám: 50% · 6p = 3p.)

Konferenciacikkek

[C1] Tamás Gábor Csapó, Géza Németh, „Transformation of irregular voice to modal voice by residual analysis and synthesis," *IEEE Signal Processing Letters*, elkészítés alatt, 2013.
(BME-PA pontszám: 0p · 100% = 0p.)

[C2] Tamás Gábor Csapó, Géza Németh, „A novel irregular voice model for HMM-based speech synthesis," *Proc. ISCA SSW8 - 8th Speech Synthesis Workshop*, (Barcelona, Spanyolország), pp. 229–234, 2013.
(BME-PA pontszám: 100% · 3p = 3p.)

[C3] Tamás Gábor Csapó, Géza Németh, „A novel codebook-based excitation model for use in speech synthesis," *IEEE CogInfoCom 2012*, (Kassa, Szlovákia), pp. 661–665, 2012.
(BME-PA pontszám: 100% · 3p = 3p.)

[C4] Tamás Gábor Csapó, Zsuzsanna Bárkányi, Tekla Etelka Gráczi, Tamás Bőhm, Steven M. Lulich, „Relation of formants and subglottal resonances in Hungarian vowels," *Proc. Interspeech 2009*, (Brighton, Egyesült Királyság), pp. 484–487, 2009.
(BME-PA pontszám: 50% · 3p = 1.5p.)

Csak kivonatban megjelent konferencia-előadások

[C5] Csapó Tamás Gábor, Németh Géza, „Irreguláris beszéd regulárissá alakítása beszédkódoláson alapuló módszerrel," *Beszédkutatás*, (Budapest), 2013. november 14–15.
(BME-PA pontszám: 100% · 0p = 0p.)

[C6] Csapó Tamás Gábor, Bárkányi Zsuzsanna, Gráczi Tekla Etelka, Beke András, Bőhm Tamás, „A magánhangzó-formánsok és a szubglottális rezonanciák összefüggése a spontán beszédben," *Beszédkutatás*, (Budapest), 2009. október 16–17.
(BME-PA pontszám: 20% · 0p = 0p.)

Egyéb, a tézispontokhoz nem kapcsolódó tudományos közlemények

Folyóiratcikkek

[J5] Tamás Gábor Csapó, Csaba Zainkó, Géza Németh, „A Study of Prosodic Variability Methods in a Corpus-Based Unit Selection Text-To-Speech System," *Infocommunications Journal*, LXV. évf., I. sz., pp. 32–37, 2010.
(BME-PA pontszám: 50% · 4p = 2p.)

[J6] Csapó Tamás Gábor, „Változatos prozódia megvalósítása szövegfelolvasó rendszerekben," *Akusztikai Szemle*, IX. évf., 3. sz., pp. 16–18, 2009.
(BME-PA pontszám: 100% · 2p = 2p.)

[J7] Csapó Tamás Gábor, Németh Géza, Fék Márk, „Szövegfelolvasó természetességének növelése," *Híradástechnika*, LXIII. évf., 5. sz., pp. 21–30, 2008.
(BME-PA pontszám: 50% · 2p = 1p.)

Konferenciacikkek

[C7] Éva Székely, Tamás Gábor Csapó, Bálint Tóth, Péter Mihajlik, Julie Carson-Berndsen „Synthesizing Expressive Speech from Amateur Audiobook Recordings," *SLT 2012*, (Miami, Florida, USA), pp. 297–302, 2012.
(BME-PA pontszám: 20% · 3p = 0.6p.)

[C8] Csapó Tamás Gábor, Németh Géza, „Prozódiai változatosság rejtett Markov-modell alapú szövegfelolvasóval," *Magyar Számítógépes Nyelvészeti Konferencia*, (Szeged), pp. 167–177, 2011.
(BME-PA pontszám: 100% · 1p = 1p.)

[C9] Tekla Etelka Gráczi, Steven M Lulich, Tamás Gábor Csapó, András Beke, „Context and speaker dependency in the relation of vowel formants and subglottal resonances - Evidence from Hungarian," *Proc. Interspeech 2011*, (Firenze, Olaszország), pp. 1901–1904, 2011.
(BME-PA pontszám: 25% · 3p = 0.75p.)

[C10] Géza Németh, Gábor Olaszy, Tamás Gábor Csapó, „Spemoticons: Text-To-Speech based emotional auditory cues," *ICAD 2011*, (Budapest), 2011.
(BME-PA pontszám: 50% · 2p = 1p.)

[C11] Csaba Zainkó, Tamás Gábor Csapó, Géza Németh, „Special Speech Synthesis for Social Network Websites," *Lecture Notes In Computer Science*, 6231: pp. 455–463, Paper 58, 2010.
(BME-PA pontszám: 50% · 6p = 3p.)

[C12] Csapó Tamás Gábor, Németh Géza, „Mássalhangzó-magánhangzó kapcsolatok automatikus osztályozása szubglottális rezonanciák alapján," *Magyar Számítógépes Nyelvészeti Konferencia*, (Szeged), 2009. december 3-4., pp. 226-237.
(BME-PA pontszám: 100% · 1p = 1p.)

[C13] Géza Németh, Márk Fék, Tamás Gábor Csapó, „Increasing Prosodic Variability of Text-To-Speech Synthesizers," *Proc. Interspeech 2007*, (Antwerpen, Belgium), pp. 474–477.
(BME-PA pontszám: 50% · 3p = 1.5p.)