Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék

Irreguláris zöngével képzett beszéd vizsgálata és modellezése

Ph.D. tézisfüzet

Bőhm Tamás Mihály
okl. mérnök-informatikus

Témavezetők:
Németh Géza, Ph.D.
Olaszy Gábor, D.Sc.

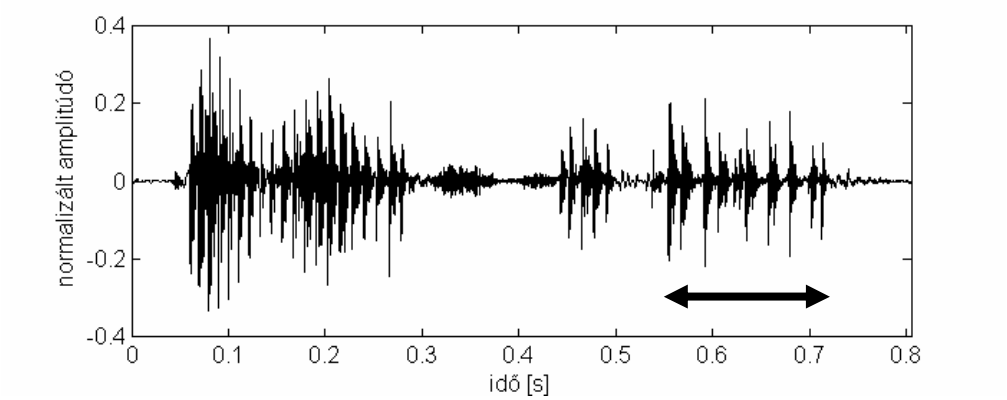
Budapest, 2009.

1. Bevezetés

Az ember régi vágya, hogy gépekkel beszédet állítson elő, vagy éppen gépekkel élőszóban kommunikáljon. Az elmúlt évtizedekben a beszédtechnológia jelentős sikereket ért el, amelyek közelebb vittek az ilyen célok valóra váltásához. Széles körben elérhetőek olyan rendszerek, amelyek bizonyos megkötésekkel az emberi beszédet megközelítő minőségben képesek szöveget felolvasni [1;2], vagy emberi beszédet képesek lejegyezni [3].

Ezekre a sikeres fejlesztésekre részben az adott lehetőséget, hogy a technológiai kutatómunka az emberi beszéd egy „idealizált" formájára összpontosított. Az idealizált beszéd itt prototípikus, fiatal felnőtt, egészséges beszélők által, grammatikus szövegek felolvasását jelenti, amelyre érvényesek a beszédtudományban általánosan elfogadott elvek (pl. a forrás-szűrő modell, vagy éppen a zönge rövidtávú periodicitása). A beszédszintetizátorokat és beszédfelismerőket többnyire az idealizált beszéd feldolgozására készítették fel, ennek megfelelő előfeltevésekkel tervezték.

Véleményem szerint azonban az idealizált beszédre tervezés korlátozza a technológia további fejlődését. A beszédjel a gyakorlatban sokszor nem idealizált beszéd. Még ha el is tekintünk az idősek és a gyermekek beszédétől, valamint a patologikus beszédtől, nagyon gyakran akkor sem teljesülnek az idealizált beszédhez kapcsolódó előfeltevések. A beszédtudomány újabb eredményei sorra mutatnak rá az ilyen alapelvek korlátaira.



1. ábra: Irreguláris fonáció egy beszédrészlet hullámformáján (nyíllal jelölt szakasz)

A hangszalagrezgésről is hagyományosan azt feltételezték a beszédtechnológiában, hogy a zöngeképzés során nagyjából periodikus, reguláris (periódusról periódusra csak kis mértékű periódusidő- és amplitúdóváltozások figyelhetőek meg). A hangszalagrezgés viszont eltérhet ettől az idealizált modelltől: például néha irregulárissá válhat, ami minőségileg eltérő, érdes, rekedtes hangzású zöngét eredményez. Bár ez, az *irreguláris fonációnak* (1. ábra) nevezett jelenség már régóta ismert, a jelenleg alkalmazott beszédtechnológiák általában nem kezelik. Ennek oka valószínűleg az, hogy korábban sokan egy normál beszédben ritka, elhanyagolható jelenségnek tekintették. Mára már azonban egyre több eredmény utal arra, hogy ez a zöngeképzési mód viszonylag gyakori és valószínűleg számos nyelvi és nem-nyelvi kommunikációs funkcióval rendelkezik a különböző nyelveken. Ilyen funkció például

a prozódiai szerkezet és az egyes érzelmi töltetek akusztikai jelölése, valamint a beszélő egyéni hangjellemzőihez való hozzájárulás.

A fentiek alapján úgy tűnik, hogy az egészséges beszédben előforduló irreguláris zöngeképzés megfelelő kezelése a beszédtechnológia számos területén hasznosítható. Például figyelembe vételével a mesterséges beszédre ültetett prozódia természetesebb hangzású lehet, vagy éppen hitelesebben kifejezhet egyes érzelmeket. Az irreguláris fonáció automatikus felismerése segíthet beszédfelismerők esetén a közlések intonációs frázisokra bontásában és egyéb prozódiai információk kinyerésében, valamint lehetővé teheti egyes érzelmi töltetek gépi felismerését. A beszélők közötti zöngeminőségbeli különbségek kihasználása pedig javíthatja a beszélőazonosító és beszélőmódosító rendszerek teljesítményét.

Ennek azonban előfeltétele, hogy rendelkezzünk az irreguláris fonáció gépi analízisét és szintézisét lehetővé tevő eljárásokkal. Disszertációmban ilyen beszédtechnológiai eljárásokat ismertetek: az I. téziscsoport egy, az irreguláris fonáció megjelenését detektálni képes algoritmusról szól, míg a II. téziscsoportban egy olyan módszert mutatok be, amely képes egy reguláris zöngével képzett beszédrészletet irregulárissá alakítani. Ezeket az eredményeket a III. téziscsoportban néhány alkalmazási lehetőséget felvető fonetikai és pszichológiai eredménnyel egészítem ki.

1.1. Irodalmi áttekintés

Irreguláris zöngeképzésnek (irreguláris fonációnak) nevezem azt a jelenséget, amikor a hangszalagok szabálytalan rezgése miatt periódusról periódusra hirtelen, nagy mértékű periódusidő- és amplitúdóváltozások láthatóak (azaz a periodicitástól való eltérés jóval meghaladja a normális jitter[1] vagy shimmer értékét [4]; 1. ábra) és ez az ép hallású személyek számára határozottan érzékelhető. Irreguláris fonációnak tekintem azt a jelenséget is, amikor az alapfrekvencia hirtelen, jelentősen, a beszélő jellemző hangterjedelme alá csökken és ezáltal érzékelhetően megváltozik a hangszínezet.

Az irreguláris zöngeképzésre a szakirodalomban általában (a fentihez hasonló) kvalitatív meghatározásokat szoktak adni [5-7] mivel még nem ismert, hogy ezt a zöngeminőséget számos akusztikai tulajdonság (pl. változások a jitter, shimmer, alapfrekvencia, intenzitás, nyitott hányad és spektrális lejtés mértékében) milyen kombinációja határozza meg.

Az irreguláris fonációt egyes források recsegő [8], érdes, rekedtes [9], nyikorgó zöngének [10], mások laringalizált, csikorgó beszédnek [11], „nyekergésfélének" [12], vagy éppen glottalizációnak nevezik [B3]. Az angol szaknyelv is számos különböző kifejezést használ a jelenségre (pl. creaky voice, vocal fry, pulsed phonation, laryngealization, glottalization). Patológiás hangszalag-elváltozások (például aszimmetrikus működés, csomó, bénulás) egyik lehetséges tünete az állandóan irreguláris hangszalagrezgés [8;13]. Munkám során azonban normális, egészséges hangszalagokkal képzett beszéddel foglalkoztam. Ezek esetén is előfordul irregularitás, de általában csak időszakosan jelentkezik és előfordulása összefüggésbe

[1] A jitter az alapfrekvencia, a shimmer az amplitúdó ingadozását jellemzi. Mivel még a reguláris zönge sem tökéletesen periodikus, az alapfrekvencia és az amplitúdó ebben az esetben is kis mértékben ingadozik.

hozható bizonyos nyelvi (szegmentális és szupraszegmentális) és nemnyelvi üzenetekkel (például az érzelmi töltettel).

Az irreguláris fonáció automatikus felismerése: A szakirodalomban fonációs típusok (reguláris vagy irreguláris) osztályozására több különböző megközelítés olvasható (például [4;7]). Ezekben közös, hogy mindegyik a vizsgált jel kisebb egységekre darabolása után egy sor akusztikus jellemzőt nyer ki a beszédjelből (lényegkiemelés), majd e jellemzők értéke alapján hoz döntést. A korábbi osztályozók számos különböző jellemzőt és döntő algoritmust alkalmaztak, de az elért pontosság (találati arány és téves riasztási arány) egyik osztályozó esetén sem elég magas ahhoz, hogy a gyakorlatban széles körben alkalmazható legyen. Az egyes jellemzők értékelésére használt korábbi módszerek (átlagok összehasonlítása, hisztogramok) vagy olyan feltételezéseken alapulnak, amelyek valószínűleg nem teljesülnek (pl. egymóduszú eloszlás), vagy nem szolgáltatnak egyetlen számszerű teljesítmény-mutatót.

Az irreguláris fonáció gépi előállítása: A szakirodalomban több olyan eljárásról olvashatunk, amelyek célja irreguláris zöngével képzett beszéd mesterséges előállítása. Ezeket az eljárásokat két csoportba sorolhatjuk: a módszerek egyik csoportja formánsszintézist (pontosabban másolás-szintézist), a másik hullámforma manipulációt (általában a jitter jelentős megnövelését) alkalmaz. Másolásszintézissel sokféle zöngeminőség (egészséges és patologikus) állítható elő [14-16], de a számos, időben változó szintézisparaméter megfelelő beállítása szakértelmet és akár több napi munkát igényelhet [16]. A hullámforma-manipulációs eljárások [17-19] pedig figyelmen kívül hagyják, hogy a jitteren kívül az irreguláris fonációnak számos egyéb akusztikai összetevője is van, amelyek szintén szerepet játszhatnak az észlelésében, továbbá jelentősen eltérő irregularitások eredményezhetnek egyforma mértékű jittert (például a pulzusok időzítésének véletlen perturbációja és azok determinisztikus mintázat szerinti változtatása).

Az irreguláris zöngeképzés előfordulása magyar beszédben: Nem ismerek olyan korábbi kutatást, amely magyar beszédben az irreguláris fonáció előfordulását vizsgálta volna. Az irreguláris fonációt említő tanulmányokban (például [9;12]) nem vizsgálták szisztematikusan annak előfordulását, legfeljebb a számítógépes beszédelemzést megnehezítő jelenségként írnak róla [10;11;20].

Az irreguláris fonáció és a beszélő személye: Több nyelven is megfigyelték, hogy az irreguláris zöngeképzés gyakorisága beszélők között eltérhet [5;6;21]. Ezen kísérletek célja azonban nem a személyfüggőség vizsgálata volt, így többnyire kevés adatközlővel dolgoztak és csak a beszélők közötti különbségeket mérték, az egy személyen belüli változékonyságot (azaz, hogy a beszélők több felvételen hasonló arányban használnak-e irreguláris fonációt) nem. Bár az irreguláris fonáció percepciójával kapcsolatban számos eredmény született, az emberek számukra ismert beszélők fonációs szokásaira vonatkozó emlékezetét tudomásom szerint korábban nem vizsgálták.

2. Kutatási célkitűzések

Célom olyan gépi eljárások kidolgozása, amelyek lehetővé teszik az irreguláris zöngeképzés által hordozott nyelvi és nemnyelvi információ kihasználását a beszédtechnológiában.

Ennek megfelelően egyik célkitűzésem egy zöngeminőség osztályozó létrehozása, amely képes magánhangzókat a korábban publikált osztályozóknál nagyobb pontossággal (magasabb találati aránnyal és alacsonyabb téves riasztási aránnyal) reguláris/irreguláris osztályokba sorolni. Megjegyzem, hogy a magánhangzók mellett a zöngés mássalhangzókat is lehetséges irreguláris fonációval képezni, azonban a gyakorlati alkalmazások szempontjából általában elegendő, ha a magánhangzók zöngeminőségéről vannak pontos információink. Ha ugyanis a prozódiai szerkezetre, a beszélő érzelmi állapotára vagy a beszélő személyére szeretnénk következtetni a zöngeminőség felhasználásával, akkor a magánhangzókra hozott döntés által biztosított szótagszintű felbontás általában elegendő.

Másik célkitűzésem egy olyan eljárás kidolgozása, amivel géppel előállítható irreguláris fonáció, azaz reguláris zöngével képzett beszéd egy kijelölt részlete irregulárissá alakítható úgy, hogy az mind akusztikailag, mind érzetileg a természetes irreguláris fonációra hasonlítson. Megjegyzem, hogy a transzformáció másik iránya (irreguláris zöngével képzett beszédrészlet regulárissá alakítása) gyakorlati szempontból kisebb jelentőséggel bír (folyamatos beszédben a reguláris fonációval ejtett szótagok vannak túlsúlyban), így ezzel nem foglalkoztam.

Célom továbbá az irreguláris fonáció és a beszélő személy kapcsolatának vizsgálata. Egyrészt a beszélő személyek közötti szisztematikus különbségek kvantitatív kimutatása egy beszédkorpusz elemzésével, másrészt annak kísérleti vizsgálata, hogy a hallgató személyek emlékeznek-e a számukra ismert beszélők fonációs szokásaira. További célkitűzés az irreguláris fonáció előfordulásának vizsgálata a magyar beszédben. Ezek a célkitűzések rámutatnak a két fentebb említett gépi eljárás (zöngeminőség osztályozó és módosító) alkalmazási lehetőségeire.

A fenti célkitűzésekhez tartozó munka nagyját angol nyelvre végeztem, de igyekeztem olyan megoldásokban gondolkodni, amelyek a magyar nyelvre való alkalmazást is lehetővé teszik.

3. Módszertan

A dolgozatomban bemutatott kutatások során a beszédtechnológiában és a kapcsolódó tudományágakban elterjedt módszerekkel dolgoztam.

A zöngeminőség osztályozó tanításához és teszteléséhez a széles körben alkalmazott TIMIT [22] beszédkorpuszt használtam. Ez lehetővé teszi eredményeim reprodukálhatóságát, valamint összehasonlíthatóságát. Ennek megfelelően az osztályozó pontosságát egy olyan korábban publikált zöngeminőség osztályozó teljesítményével hasonlítottam össze, amely ugyanezt a tanító- és teszthalmazt használta. A döntést szupport vektor géppel (Support Vector Machine, SVM)

végeztem, amelyet egy nyilvánosan elérhető Matlab eszköztárral (OSU SVM[2]) valósítottam meg.

A zöngeminőség módosító eljárást általánosan elfogadott módszerekkel értékeltem akusztikai és percepciós szempontból. Az akusztikai vizsgálat során Holmberg és szerzőtársai, valamint Hanson módszerével [23;24], a felharmonikusok amplitúdó viszonyai alapján következtettem a glottális forrásjel egyes spektrális jellemzőire. Ez a módszer ugyanis lehetővé teszi, hogy az irreguláris zöngével képzett beszédre gyakran pontatlan inverz szűrés nélkül végezzünk méréseket a glottális forrásjellel kapcsolatban. A percepciós tesztben a kísérleti személyek az elterjedt ötfokozatú skálán értékelték a hangingereket. Mindkét értékelés során az eredményeket a természetes beszédben előforduló irreguláris fonációra kapott eredményekkel hasonlítottam össze.

4. Új eredmények

4.1. I. téziscsoport: Eljárás reguláris és irreguláris zöngével képzett magánhangzók automatikus osztályozására

A szakirodalomban több különböző zöngeminőség osztályozóról olvashatunk. Az ezek által használt akusztikus jellemzők között nagyon kevés az átfedés és a különböző jellemzők valószínűleg az irreguláris fonáció más-más sajátosságait ragadják meg. Így az egyes cikkekben leírt akusztikus jellemzők egy zöngeminőség osztályozó rendszerbe integrálásával teljesítményjavulás érhető el. Egy ilyen osztályozó rendszert vezet be ez a téziscsoport.

A rendszer célja, hogy magánhangzókról reguláris/irreguláris döntést hozzon. Munkám során a TIMIT amerikai angol beszédkorpuszból [22] az 1. és a 2. nyelvjárási régióhoz (rendre New England és Northern) tartozó beszélők felvételeit használtam. Ezekhez a felvételekhez rendelkezésemre álltak zöngeminőség címkék (azaz az osztályozó kívánt kimenete az adott magánhangzóra)[3], amelyek lehetővé tették felügyelt tanítási módszerek alkalmazását. A felvételeket a TIMIT készítői által javasolt módon osztottam fel tanító- és teszthalmazra.

I.1. tézis: [C2] *Újszerű elemzési módszert (a más területeken elterjedt ROC görbéket) alkalmaztam zöngeminőség osztályozókban használt akusztikus jellemzők értékelésére. Az ROC elemzés lehetővé teszi az egyes jellemzők reguláris-irreguláris szeparációs képességének a választott küszöbtől független vizsgálatát és számszerűsítését, a jellemzők eloszlásával kapcsolatos előfeltevések nélkül.*
Az ROC elemzést a szakirodalomban elérhető [4;7] alábbi hat akusztikus jellemző értékelésére használtam fel:

- *Alapfrekvencia (F0)*
- *Normalizált RMS intenzitás (Normalized RMS intensity, NRMS)*
- *Simított energiakülönbség (Smoothed Energy Difference, SED)*
- *Energiacsúcs emelkedés és ereszkedés (PoWer Peak falling and rising, PWP)*

[2] http://sourceforge.net/projects/svm
[3] A zöngeminőség címkéket Kushan Surana (MIT) készítette [4] és bocsátotta rendelkezésemre, amit ezúton is szeretnék megköszönni.

- *Eltoláskülönbség amplitúdó (Shift-Difference amplitude, SD)*
- *Kereten belüli periodicitás (IntraFrame Periodicity, IFP)*

A fenti jellemzőkből öt olyan jellemzőt dolgoztam ki (algoritmikus változtatásokkal és konstansok finomhangolásával), amelyek átlagosan pontosabban képesek szétválasztani a kétféle zöngeminőséget, mint az ezen jellemzők korábban publikált változatai. Az átlagos pontosságot az ROC görbe alatti területtel mértem.

Zöngeminőség osztályozók által használt akusztikus jellemzők szeparáló képességét a más területeken jól ismert Receiver Operating Characteristic (ROC) görbék segítségével mértem, amely – szemben korábban alkalmazott módszerekkel (hisztogramok, átlagok összehasonlítása) – nem tesz előfeltevéseket a jellemző értékeinek eloszlásával kapcsolatban. Az ROC görbe alatti terület (Area Under Curve, AUC) a különböző küszöbértékekhez tartozó átlagos osztályozóképességet mutatja, így az AUC a küszöbértéktől függetlenül képes számszerűen értékelni az akusztikus jellemzők teljesítményét [25].

Változatlan formában újraimplementáltam a Surana [4] által publikált négy akusztikus jellemzőt (F0, NRMS, SED és SD) és az Ishi és szerzőtársai [7] által leírt három jellemző közül kettőt (PWP és IFP; a harmadik jellemző a zöngés és zöngétlen beszédrészletek elkülönítésére szolgál, így a kitűzött feladat szempontjából nem releváns). Majd két különböző módon javítottam az akusztikus jellemzők reguláris-irreguláris szétválasztási képességén. Egyrészt az algoritmusok által használt konstansok értékét szisztematikus tesztekkel állítottam be. Másrészt magukon az algoritmusokon változtattam és különböző alternatívákat vizsgáltam meg. Mindezek mellett a keretezést alkalmazó ismertetőjegyek esetén az időfelbontás javítása érdekében áttértem az átfedő keretek (30 ms hosszú keretek, 5 ms léptetéssel) alkalmazására. Az eredeti jellemzők számításának alapötletét és a továbbfejlesztéseket az alábbiakban ismertetem (részletes ismertetésük disszertációmban olvasható):

Alapfrekvencia (F0): Egy hagyományos autokorreláció alapú alapfrekvencia-meghatározótól azt várjuk, hogy egy irreguláris fonációval képzett beszédrészlet esetén a reguláris beszédnél alacsonyabb F0 értéket határoz meg, vagy zöngétlennek tekinti a beszédrészletet (amit 0 Hz-es visszatérőértékkel jelez) [4].
A normalizált autokorreláció a véges kerethossz miatt egy lineárisan ereszkedő burkolóval rendelkezik, amely a maximális eltolásnál éri el a nullát. Hogy így az elemzésből ne zárjam ki az alacsony alapfrekvenciához tartozó csúcsokat, normalizált autokorreláció helyett torzítatlan autokorrelációt használtam. A zöngésségi küszöb értékét 0,35-re állítottam, mert a 0,25-0,50 tartományban végzett szisztematikus tesztek alapján ez magasabb teljesítményt nyújt, mint az eredeti 0,46-os érték. Továbbá a csúcsválasztás kritériumai közül eltávolítottam a harmadikat (amelyik az egymástól nagyjából egyenlő távolságra levő csúcsok esetén teljesül), mert ez ebben az algoritmusban számos felezési hibát eredményezett. A változtatások eredményeképpen az ROC görbe alatti terület 0,87-ről 0,93-ra emelkedett.

Normalizált RMS intenzitás (Normalized RMS, NRMS): Egy egész mondatra normalizált RMS intenzitás mértéktől azt várhatjuk, hogy alacsonyabb az irreguláris zöngével képzett beszéd esetén, mint a regulárisban (az előbbi esetben ugyanis a hangszalagok általában ritkábban csapódnak össze, mint reguláris beszédben) [4].
A keretek intenzitásának teljes mondatra történő normalizálása megtévesztő lehet: a mondat során ugyanis jelentős intenzitásváltozások lehetnek. Ezért a teljes mondat

helyett csak a vizsgált magánhangzóra és környezetére normalizálok. A környezet méretére vonatkozó vizsgálatok azt mutatták, hogy a magánhangzóra és ±25 ms környezetére normalizálás enyhén növeli az ROC görbe alatti területet (0,84-ről 0,86-ra).

Simított energiakülönbség (Smoothed energy difference, SED): Ez a jellemző az irreguláris fonáció esetén a glottális impulzusok nagyobb időbeli távolsága miatt látható gyors energiaváltozásokat használja ki. Ehhez egy finom időfelbontású energiamenet kétféle (6 és egy 16 ms hosszú ablakkal) simított változatának a maximális különbségét veszi [4].
A két simítóablak méretének szisztematikus vizsgálatával arra jutottam, hogy a 2 ms és 4 ms hosszú ablakok javítják legnagyobb mértékben a reguláris és irreguláris magánhangzók szétválasztását. Ezenfelül a két simított jel különbségének maximuma helyett az abszolút maximum számítása kiküszöböl néhány hibás mérést. A két módosítás eredményeképpen az ROC görbe alatti terület 0,08-dal emelkedett (0,74-ről 0,82-re).

Energiacsúcs emelkedés és ereszkedés (Power peak rising and falling, PWP): Úgy, mint az SED, ez a jellemző is az irreguláris glottális impulzusok körül látható gyors energiaváltozásokon alapul. Egy „nagyon rövid távú energiakontúr" csúcsait keresi meg, majd a csúcs előtt és után meghatározza az emelkedés/ereszkedés mértékét [7].
Az algoritmus által használt két konstanst (az energiamenet csúcsainak megtalálásánál figyelembe vett korábbi és későbbi pont távolsága a vizsgált ponttól, valamint az energia emelkedésének és csökkenésének kiszámításánál felhasznált környezet mérete) szisztematikus tesztekkel finomhangolva (rendre 3 helyett 4-es és 5 helyett 4-es értékre) javult a jellemző teljesítménye: a ROC görbe alatti terület 0,79-ről 0,85-re emelkedett.

Eltoláskülönbség amplitúdó (Shift-difference amplitude, SD): Ez a jellemző a beszéd harmonikus-zaj arányát becsli. A különböző eltolásokkal számolt különbségnégyzet jelek minimumát veszi és normalizálja, majd idő szerint átlagolja [4].
A Kochanski és szerzőtársai [26] által publikált „aperiodicitási mérték" egy alternatív módszer az eltoláskülönbség kiszámítására (az eredeti SD jellemzőt is ez inspirálta). Ez az algoritmus egy olyan függvényt állít elő, amely megadja a harmonikus-zaj arány időbeli lefutását a magánhangzóban. Ennek a függvénynek az átlaga a magánhangzó továbbfejlesztett SD jellemzője. Ez az alternatív algoritmus 0,75-ről 0,88-ra növeli az ROC görbe alatti területet.

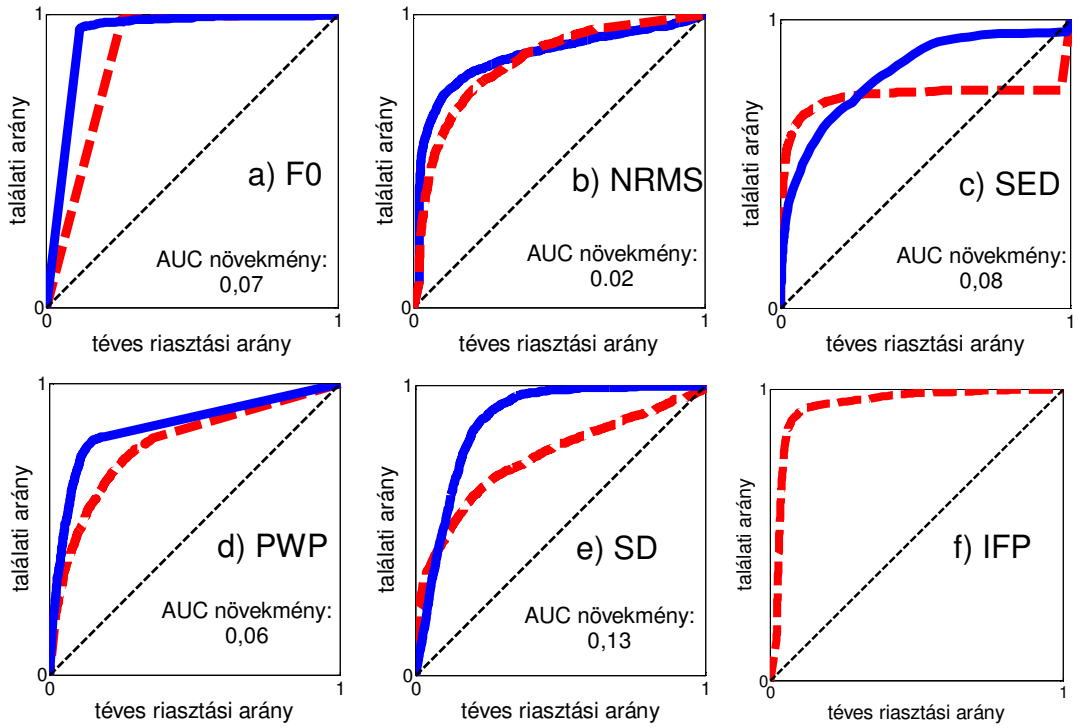
Kereten belüli periodicitás (IFP): Hasonlóan az SD-hez, ez a jellemző is a harmonikus-zaj arányt méri. A torzítatlan autokorreláció függvényen meghatározza az első prominens csúcsot, majd az ennek egész számú többszöröseinél vett autokorreláció értékek minimuma lesz a kereten belüli periodicitás mértéke [7].
Az IFP jellemző már eredeti formájában is kiválóan osztályoz, így ezt nem fejlesztettem tovább.

Az akusztikus jellemzők értékelése során kiszámítottam a jellemzők értékét a tanítóhalmaz összes magánhangzójára, majd a zöngeminőség címkék segítségével kirajzoltam a jellemzők ROC görbéit és kiszámítottam a görbe alatti területet (AUC).

A módosítások számszerű értékelésére az ROC görbe alatti terület változását használtam: ha növekedett a terület, akkor átlagosan javult a jellemző osztályozási teljesítménye (2. ábra, folytonos vonalak).

Megállapítható, hogy mind a hat korábban publikált jellemző valamilyen szinten képes szétválasztani a reguláris és az irreguláris zöngével képzett magánhangzókat (a ROC görbe nagyja a találgatásnak megfelelő átló felett van és ennek megfelelően az AUC értékek mind nagyobbak 0,5-nél). A legtöbb találatot és legalacsonyabb téves riasztási arányt az IFP-vel lehet elérni, amelyet az F0 jellemző követ. Az SED-hez tartozó ROC görbe egy rövidebb szakasza az átló alatt fut, azaz csak ezen jellemző alapján, bizonyos küszöbértékek esetén többnyire téves döntéseket hozna egy zöngeminőség osztályozó rendszer, azonban az esetek többségében helyesen döntene.

Amint a 2. ábrán látható, a módosított jellemzők ROC görbéi többnyire az eredeti jellemzők görbéi felett futnak. Ez arra utal, hogy a módosított jellemzők átlagosan pontosabban képesek szétválasztani a reguláris és irreguláris zöngével képzett magánhangzókat. Az ROC görbe alatti terület növekedése számszerűen alátámasztja, hogy a hat akusztikus jellemzőből öt esetén javasolt módosítások teljesítményjavulást eredményeztek.



2. ábra: A hat akusztikus jellemző ROC görbéje (a tanítóhalmazon számolva). A szaggatott vonalak az irodalomban leírt jellemzőkhöz, a folytonos vonalak a továbbfejlesztett jellemzőkhöz tartoznak.

I.2. tézis: [C2] *Létrehoztam egy olyan automatikus zöngeminőség osztályozót, amely a korábbi, összehasonlítható rendszernél [4] pontosabban sorol magánhangzókat reguláris/irreguláris osztályokba. Az osztályozó rendszerbe foglalja az I.1. tézisben ismertetett és továbbfejlesztett akusztikai jellemzőket és a döntést SVM segítségével hozza. Így kimutattam, hogy több korábbi rendszer akusztikus jellemzőinek együttes alkalmazásával és továbbfejlesztésével jelentős pontosságnövekedés érhető el.*

Az SVM-alapú zöngeminőség osztályozó a hat akusztikus jellemző (öt esetben a továbbfejlesztett változat) alapján dolgozik. Az alkalmazott RBF (Radial Basis Function) kernellel rendelkező SVM betanítását két paraméter szabályozza: C, a helytelen osztályozás költsége és γ, a Gauss kernel jellemzője. Először kimerítő kereséssel meghatároztam a két paraméter értékét. Ehhez a tanítóhalmaz egy részét használtam (1380 reguláris és ugyanannyi irreguláris magánhangzót). Minden vizsgált paraméter-kombinációra háromszoros és tízszeres keresztvalidációt végeztem és átlagoltam a kapott két találati arányt és két téves riasztási arányt. A két átlagolt teljesítményjellemzőt ezután azonos súllyal vettem figyelembe az egyes paraméterkonfigurációk pontosságának meghatározásához. A legmagasabb pontosságot a C=0.0313 és γ=0.0313 értékekkel értem el.

A fenti paraméterekkel és a tanítóhalmaz egy kiegyensúlyozott (azaz ugyanannyi reguláris és irreguláris elemet tartalmazó) részhalmazával betanítottam az SVM-et. A kiegyensúlyozott részhalmaz tartalmazta mind az 1403 irreguláris magánhangzót, valamint ugyanennyi, véletlenszerűen kiválasztott reguláris magánhangzót.

A teszthalmazon a zöngeminőség osztályozó 98,85%-os találati arányt és 3,47%-os téves riasztási arányt ért el. Az eredményeket a korábbi zöngeminőség osztályozók közül Surana rendszerével [4] tudom összehasonlítani, mert ez szintén magánhangzókon hozott hang-szintű reguláris/irreguláris döntést. Surana ugyanazon a teszthalmazon 91,25%-os találati és 4,98%-os téves riasztási arányt közölt. Tehát a jelen tézisben javasolt osztályozó a korábbi, azzal összehasonlítható rendszerhez képest 7,60 százalékponttal több találatot és 1,51 százalékponttal kevesebb téves riasztást produkált. A bemutatott rendszer teljesítménye valószínűleg megfelelő számos gyakorlati alkalmazáshoz.

4.2. II. téziscsoport: Eljárás reguláris zöngével képzett beszéd irregulárissá alakítására

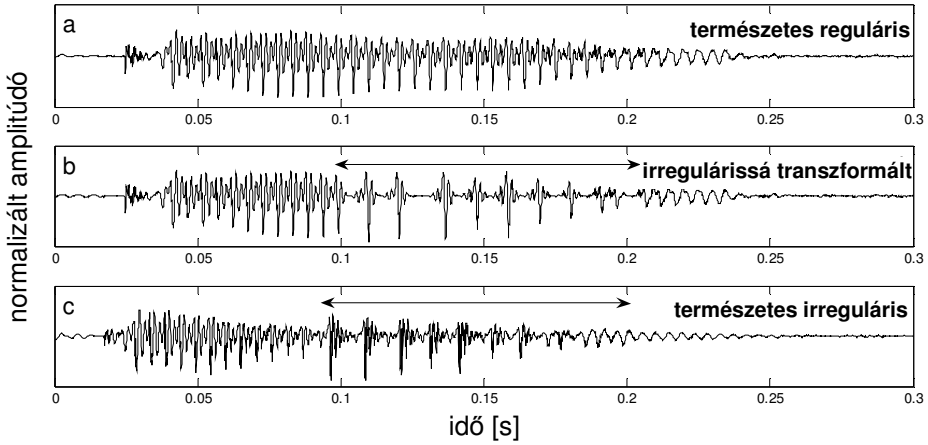
Ez a téziscsoport egy olyan eljárást vezet be, amely reguláris zöngével képzett beszédrészleteket képes irregulárissá alakítani. Az eljárás az egyes alapperiódusok amplitúdójának külön-külön történő skálázásán alapul. A transzformált beszédet értékeltem mind érzeti, mind akusztikai szempontból, természetes (reguláris és irreguláris zöngével képzett) beszéddel összehasonlítva.

II.1. tézis: [C3, C4, J1] *Eljárást dolgoztam ki reguláris zöngével képzett beszéd irregulárissá alakítására, amely az irregularitást az egyes alapperiódusok amplitúdójának egyenkénti skálázásával (akár nulla szintre) hozza létre. A skálázó faktorok félautomatikus beállítására az „impulzusmintázat-másolás” módszert dolgoztam ki.*

Az eljárás zöngeszinkron ablakozással megközelítőleg szétválasztja az alapperiódusokat, majd az egyes periódusok amplitúdóját egyenként skálázza egy periódusonként beállított faktorral (gyakran nulla szintre, azaz egyes periódusokat kitörölve a jelből), végül átlapolt összeadással (overlap-and-add) előállítja a kimenő beszédjelet. Bemenete a beszédjel ($x[n]$) és a glottális impulzusok hozzávetőleges időpontjai (*pitchmark*-ok; p_i, $1 \leq i \leq N$) a transzformálni kívánt régióban. Az első lépés az egyes alapperiódusok „szétválasztása" a glottális impulzusok környezetének Hanning-ablakolásával. Az ablak ($w_i[n]$) csúcspontját a pitchmarkon éri el és az előző pitchmarktól a következő pitchmarkig tart (tehát két alapperiódust fed le és aszimmetrikus is lehet). Ez az ablakozási eljárás – amely megegyezik a PSOLA (Pitch Synchronous OverLap-and-Add) algoritmus [27] analízis szakaszával – az egyes alapperiódusok egy durva közelítését nyeri ki különálló jelekbe. Ezután minden egyes jelet mintánként beszorzunk egy kézzel beállított skálázó faktorral (s_i), majd az egyes jeleket átfedve és összeadva (overlap-and-add) újraszintetizáljuk a teljes beszédjelet:

$$\hat{x}[n] = \sum_{i=1}^{N} \left(s_i x[n] w_i[n] \right)$$

ahol $\hat{x}[n]$ a kimenő beszédjel. A skálázó faktorok az egyes periódusokat erősíthetik (s_i>1), csillapíthatják (s_i<1), kitörölhetik (s_i=0) vagy módosítás nélkül meghagyhatják (s_i=1). Egy transzformált beszédfelvétel látható a 3.b. ábrán.



3. ábra: Egy reguláris zöngével végződő beszédrészlet (a) és annak transzformált változata (b), valamint egy természetes irreguláris beszédrészlet (c). A vízszintes nyilak az irreguláris szakaszokat jelölik.

Egy-egy alapperiódus csillapítása vagy éppen nullázása a periódusban jelen lévő háttérzaj szintjét is lecsökkenti (például több egymás utáni periódus törlésekor a zaj hiánya a természetesség érzetét csökkentheti). Ennek az elkerülésére az eljárás háttérzajt (b[n]) ad a csillapított és nullázott periódusokhoz. A zajt például a felvétel végéről lehet kivágni, majd ablakozás után ($1 - s_i$)-vel skálázva kompenzálja az eredeti periódus skálázásából adódó zajenergia-veszteséget:

$$\hat{x}[n] = \sum_{i=1}^{P} \left(s_i x[n] w_i[n] + \max(1 - s_i, 0) b[n - p_i] w_i[n] \right)$$

A skálázó faktorok félautomatikusan beállíthatók „impulzusmintázat-másolással". Ennek lényege, hogy a kialakítandó impulzusmintázatot (a glottális impulzusok időbeli távolságai és amplitúdói) egy természetes beszédben előforduló irreguláris régió alapján modellezzük – azaz egy modell felvételt használunk, amelynek impulzusmintázatát a javasolt módszerrel átmásoljuk a transzformálandó beszédrészletre. Az impulzusmintázat kezdetben egy olyan vektor, amely a mintaként szolgáló irreguláris beszédrészlet glottális impulzusainak relatív amplitúdóját tartalmazza. Ha egy irreguláris alapperiódus jelentősen hosszabb, mint a $T0_{ref}$ referencia periódusidő (pl. kétszer vagy háromszor olyan hosszú, mint a referencia, amit néhány megelőző periódus átlagaként számol), akkor az eljárás megfelelő számú nullát szúr be az impulzusmintázatba – ezeken a helyeken ugyanis egy vagy több periódust törölni kell a jelből.

II.2. tézis: [J1, C3, C4] *Percepciós kísérletet dolgoztam ki, amellyel igazoltam, hogy a II.1. tézis eljárásával transzformált felvételeket a kísérleti személyek az irreguláris fonációhoz hasonló érdességűnek és a manipulálatlan beszédet megközelítő természetességűnek ítélik meg.*

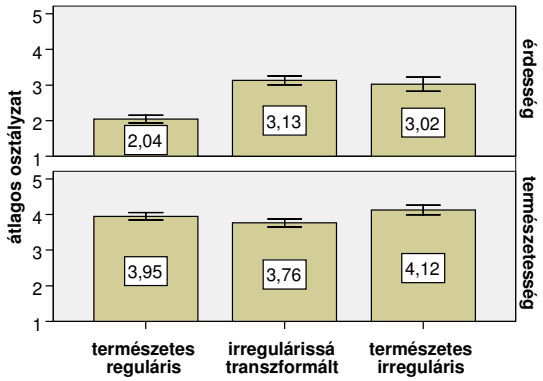
A II.1. tézisben leírt eljárás gyakorlati alkalmazhatóságának egyik fontos feltétele az, hogy elfogadják-e humán hallgatók. Ezt egy meghallgatásos kísérletben mértem, amellyel az volt a célom, hogy értékeljem az eljárás által a beszédjelben előidézett észlelt érdességet és az esetleges természetesség-romlást. Mivel az irreguláris fonáció által keltett észleletet általában érdes hangként írják le, az észlelt érdesség mérése lehetővé teszi, hogy arra következtessek, vajon a transzformált felvételek az irreguláris fonációhoz hasonlóan hangzanak-e. Az észlelt érdességtől függetlenül a természetesség esetleges csökkenése arra hívná fel a figyelmet, hogy az eljárás hallható műtermékeket hozott létre. Az eljárás érzeti értékelése során a közlésvégi irreguláris fonációra összpontosítottam. A transzformált beszédrészletekre kapott eredményeket olyan manipulálatlan, természetes beszédközlésekkel hasonlítottam össze, amelyek egy része reguláris, más része irreguláris zöngével végződött.

Az angol nyelven végzett percepciós kísérlet két tesztből állt. Az egyikben a kísérleti személyek (12 amerikai angol anyanyelvű) a beszédingerek természetességét értékelték, a másikban pedig ugyanazoknak az ingereknek az érdességét ítélték meg. A válaszokat mindkét tesztben egy ötpontos skálán adták.

Az eredmények az 4. ábrán láthatóak. Amikor egy reguláris közlésvégi fonációval ejtett természetes felvétel végét irregulárissá alakítottam a javasolt eljárással, az érdességítéletek 1,09 ponttal emelkedtek ($p<0.0005$), de a természetesség osztályzatokban ez a transzformáció csak nemszignifikáns, 0,19 pontos csökkenést eredményezett ($p=0.11$, nem szignifikáns). A transzformáció nem csak jelentősen megemelte az észlelt érdesség fokát, hanem az irreguláris fonációval ejtett természetes beszéd érdességének megfelelő szintre növelte: a különbség az ítéletek átlagában mindössze 0,11 pont volt ($p=0.91$, n. sz.). Tehát a kísérleti személyek észlelték a transzformált felvételek megnövekedett érdességét, természetesnek vélték azokat és

nem hallottak különbséget az érdesség mértékében az eredetileg irreguláris és az irregulárissá transzformált ingerek között.

Következmény: *A korábbi eljárásoktól eltérő módon a II.1. tézis eljárása az alapperiódusok időzítését és amplitúdóját hirtelen, nagy mértékben változtatja meg (például megkétszerezi a periódusidőt egy periódus kitörlésével). A meghallgatásos kísérlet eredményei azt mutatják, hogy az egymás utáni glottális ciklusok időtartamának és amplitúdójának hirtelen, nagy mértékű változása képes kiváltani az irreguláris fonációnak megfelelő érdesség észleletét.*

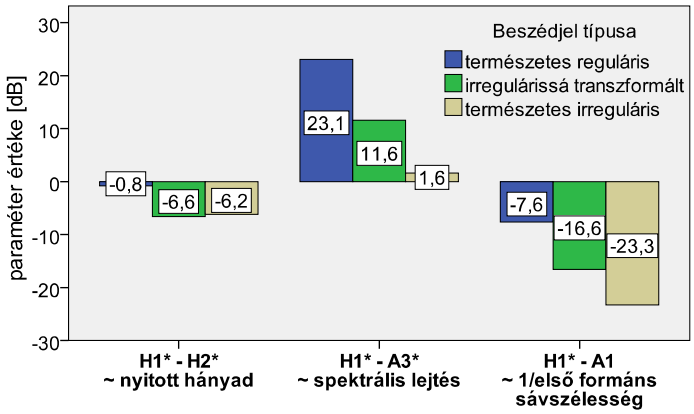


4. ábra. A meghallgatásos kiértékelés során kapott átlagos osztályzatok. A függőleges szakaszok a 95%-os konfidencia-intervallumokat jelölik.

II.3. tézis: [C3] *Ismert vizsgálati módszert átvéve igazoltam, hogy a II.1. tézis eljárása a beszéd több releváns akusztikai paraméterét (spektrális lejtés, nyitott hányad, első formáns sávszélessége) az irreguláris fonációra jellemző értékek irányában módosítja.*

Az irreguláris fonációt – regulárissal összehasonlítva – általában öt akusztikai tulajdonsággal jellemzik (a részletes magyarázatok [6;14]-ben olvashatóak): (1) az F0 alacsonyabb és a jitter magasabb; (2) a teljes intenzitás alacsonyabb; (3) a nyitott hányad (open quotient, OQ) alacsonyabb; (4) az első formáns sávszélessége (B1) nagyobb; (5) a spektrum lejtése (spectral tilt, TL) alacsonyabb. Könnyen belátható, hogy a II.1. tézisben leírt transzformáció – mivel egyes alapperiódusok jelének eltávolításával jár – jelentősen megnöveli és perturbálja a glottális impulzusok távolságát és a teljes intenzitásszintet is csökkenti.

A transzformáció hatását a (3)-(5) akusztikai jellemzőre mérésekkel vizsgáltam. A méréseket [23] szerint a frekvenciatartományban végeztem ([24] szerinti korrekciókkal): az OQ-t az első harmonikus amplitúdójával a második harmonikus amplitúdójához viszonyítva (azaz a H1*-H2* különbséggel dB-ben); a TL-t az első harmonikus amplitúdójával a harmadik formánsamplitúdóhoz viszonyítva (H1-A3*); a B1-et az első harmonikus amplitúdójával az első formánsamplitúdóhoz viszonyítva (H1*-A1). A méréseket a II.2. tézis meghallgatásos kísérletében felhasznált hanganyagon végeztem el, amely tartalmazott eredetileg irreguláris és reguláris zöngével végződő közléseket is, valamint az utóbbiakat a javasolt módszerrel irregulárissá transzformálva. A mérési eljárás részletei disszertációmban olvashatóak.



5. ábra: Az akusztikai értékelés során mért három spektrális paraméter átlagai (dB-ben)

Az eredmények a 5. ábrán láthatóak. A transzformált felvételeken mért H1*-H2* szignifikánsan alacsonyabb volt, mint az eredeti felvételeken ($p<0.0005$) és megközelítőleg azonos volt a természetes irreguláris beszédben mért átlagos értékkel ($p=0.97$, n.sz.). Tehát a nyitott hányad szempontjából a transzformált beszédrészletek nem különböztek a természetes irreguláris zöngeképzéssel ejtett beszédrészletektől. A H1*-A3* (és ennek megfelelően a spektrális lejtés) is szignifikánsan csökkent a transzformáció hatására ($p=0.001$), az irreguláris beszédre számolt középérték felé elmozdulva, de attól még szignifikánsan eltérve ($p=0.033$). Ugyanezt mondhatjuk el a B1-nek megfelelő H1*-A1 értékekről: ezek szintén szignifikánsan alacsonyabbak voltak a transzformáció után, mint előtte ($p<0.0005$). Bár az átlag közelebb van a természetes irreguláris beszédrészletek H1*-A1 átlágához, mint a természetes regulárisokéhoz, attól még szignifikánsan eltér ($p=0.001$).

Összefoglalva, az F0 és a teljes intenzitás csökkentése, valamint a jitter növelése mellett a II.1. tézisben javasolt eljárás az irreguláris fonáció néhány további akusztikus korrelátumát is reprodukálja. Ezért az eljárás jelentős előrelépést jelent a korábbi eljárásokhoz képest, amelyek csak a jitter és a shimmer manipulációjára összpontosítottak.

4.3. III. téziscsoport: Irreguláris fonáció előfordulásának gyakorisága és a beszélő személye

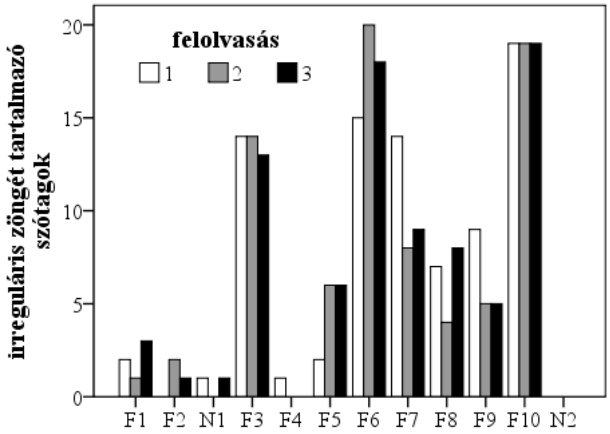
A téziscsoport egyrészt egy olyan tanulmány eredményeit mutatja be, amely magyar beszédben az irreguláris zönge előfordulási gyakoriságát általánosságban, valamint beszélő személyenként elemezte. Másrészt két angol nyelvre végzett kísérletet ismertet, amelyek azt vizsgálták, hogy a beszélő jellemző közlésvégi zöngeminőségére (ismert beszélők esetén) emlékeznek-e a kísérleti személyek illetve hogy (ismeretlen beszélők esetén) rövid tanulás során meg tudják-e jegyezni ezt az egyéni jellemzőt.

III.1. tézis: [B1] *Kísérleti eljárást dolgoztam ki, amellyel kimutattam, hogy az irreguláris zönge előfordul a normál magyar beszédben, és hogy a mondatvégi irreguláris fonáció előfordulási aránya a beszélő személyek között jelentősen eltérhet, ugyanakkor (a vizsgált beszélők esetén) személyen belül csak kisebb eltéréseket mutat.*

Vizsgálataimhoz 12 személy (2 nő, 10 férfi, életkoruk 23–33 év, kivéve egy férfi beszélőt (F3), aki 64 éves) beszédét rögzítettem. Minden adatközlő felolvasott egy 11 mondatból (7 kijelentő, 3 kérdő, 1 felkiáltó) álló szöveget. A felolvasást rövid szünet után még kétszer megismételték.

A felvételeken bejelöltem az irreguláris zöngével képzett beszédszakaszokat. A következetes címkézés érdekében konszenzusos eljárást alkalmaztam: a korpusz összes felvételén rajtam kívül egy másik címkéző személy is egy rögzített kritériumrendszer alapján megjelölte az irreguláris beszédszakaszokat, majd a párhuzamos címkeállományokat összevetettük (az egyesítés módját disszertációmban ismertetem).

Az összes szótag 9%-ában fordult elő irreguláris fonáció. Ez meglepően magas arány, ha figyelembe vesszük, hogy egy eddig többnyire figyelmen kívül hagyott jelenségről van szó. Eredményeim tehát azt mutatják, hogy az irreguláris zöngeképzés nem ritka jelenség magyar beszédben: az általam vizsgált beszédkorpuszban átlagosan minden 11. szótagban előfordult ez a fonációs típus.



6. ábra: Irreguláris zöngét tartalmazó mondatvégi (utolsó vagy utolsó előtti) szótagok száma személyenként (a három felolvasásra bontva)[4]

Mondatvéginek tekintettem azokat a címkéket, amelyek érintik egy mondat utolsó vagy utolsó előtti szótagját. A 6. ábrán látható, hogy a mondatvégi irreguláris fonáció előfordulási aránya nagymértékben eltérő az egyes adatközlők esetén. Három beszélő esetén gyakran, öt esetén pedig ritkán fordult elő ez a zöngeképzési mód. A beszélők a három felolvasás során hasonló arányban képeztek irreguláris zöngét: 7 személy esetén a három felolvasás nagyjából ugyanannyi irreguláris szótagot tartalmaz, míg 5

[4] N2 beszélő egyik felolvasásában sem találtam mondatvégi irreguláris zöngét.

személy (F5, F6, F7, F8 és F9) esetén láthatóan eltérések vannak, de ezek kis mértékűek.

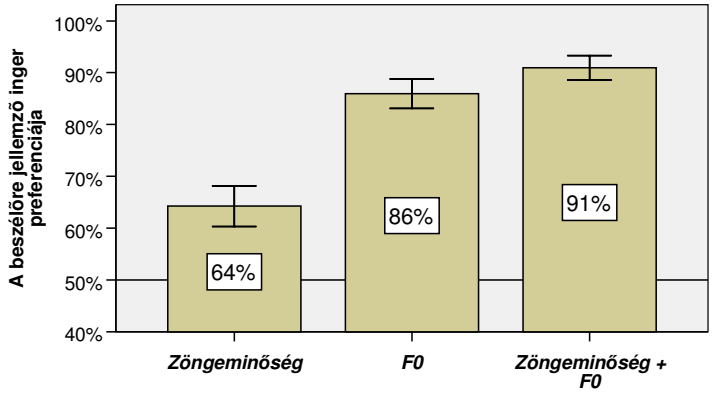
Az eredmények azt mutatták, hogy, összhangban a korábbi informális megfigyelésekkel, az irreguláris fonáció gyakorisága valószínűleg jellemző az egyes beszélőkre, különösen a mondatok végén: ugyanannak a szövegnek a három felolvasását összehasonlítva, az egyes adatközlők mindhárom alkalommal a kiejtett szótagok körülbelül ugyanakkora hányadát képezték irreguláris zöngével.

Következmény: *Valószínűsíthető, hogy jelentős minőségjavulás érhető el a magyar nyelven működő beszédtechnológiákban, ha azokat felkészítik az irreguláris fonáció kezelésére. Valószínűsíthető továbbá, hogy beszélőfelismerő rendszerek jelenlegi akusztikus jellemzőkészletét kiegészítve a közlésvégi irreguláris fonáció előfodulásának gyakoriságával teljesítményjavulás érhető el. Ez a gyakorisági mérték hasznos lehet továbbá az igazságügyi hangazonosításban is.*

III.2. tézis: [J1, B3, C5, C8] *Angol nyelvre végzett kísérlettel kimutattam, hogy a kísérleti személyek emlékeznek a számukra jól ismert beszélők közlésvégi fonációs szokásaira (irreguláris zönge előfordulási gyakoriságára).*

A III.1. tézis szerint szisztematikus különbségek mutathatók ki egyes beszélők között a mondatvégi – így egyben a közlésvégi – irreguláris fonáció előfordulásának gyakoriságában. Ebben a tézispontban azt vizsgáltam, hogy a hallgató személyek vajon emlékeznek-e számukra jól ismert beszélők hangjának e jellemzőjére és hozzá tudnak-e férni ehhez az információhoz.

Az ingeranyagot négy beszélő – egy férfi és egy nő, akik a közlések végén gyakran használtak irreguláris zöngét, valamint egy másik férfi és egy másik nő, akik a közlések végén szinte mindig reguláris fonációt alkalmaztak – négy-négy rövid felvétele alapján állítottam össze (ezeken mindegyik beszélő minden alkalommal a rá jellemző zöngeminőséggel zárta a közlését). A 9 hallgató személyesen ismerte a beszélőket. Az összesen 16 felvételből három-három manipulált változatot készítettem: az egyiken a zöngeminőséget, a másikon az átlagos alapfrekvenciát, míg a harmadikon mindkettőt módosítottam. A zöngeminőség módosításához a II.1. tézis eljárását alkalmaztam. A kísérlet során minimál párokat mutattam be a hallgatóknak (amelyek természetességét meghallgatásos tesztben ellenőriztem): egy felvétel eredetijét és ugyanannak a felvételnek az egyik manipulált változatát, véletlenszerű sorrendben. Így a párok két tagja között kizárólag (a) a végső zöngeminőségben, (b) az alapfrekvencia középértékében vagy (c) a végső zöngeminőségben és az alapfrekvencia középértékében volt különbség. A kísérleti személyek feladata az volt, hogy megállapítsák, az elhangzott hanginger-pár melyik tagja felel meg a beszélő hangjának.



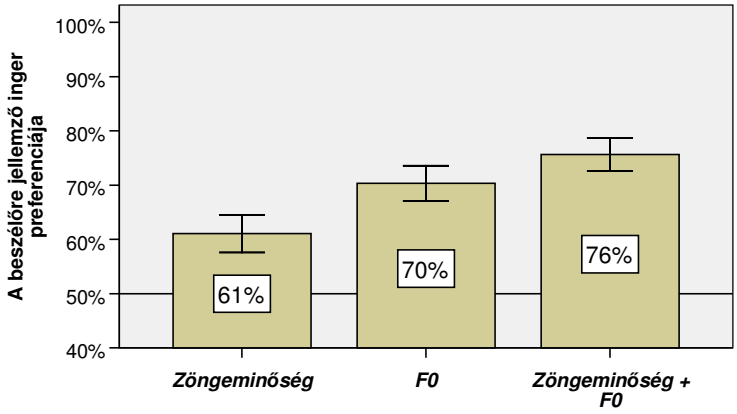
7. ábra: A beszélők szokásos zöngeminőségével, F0 átlagával, vagy mindkettővel rendelkező ingerre adott válaszok aránya (a hallgatók ismerték a beszélők hangját). A vízszintes vonal az 50%-os találati aránynak felel meg, míg a függőleges szakaszok az átlaghoz tartozó konfidencia-intervallumokat jelölik.

Az eredményeket a 7. ábra foglalja össze, amely a három feltételhez tartozó találati arányokat mutatja („találatnak" tekintem azokat a próbákat, amelyekben a kísérleti személy az eredeti, manipulálatlan beszédmintát preferálta egy manipulálttal szemben). A zöngeminőség feltétel szolgál a kísérlet hipotézisének (azaz, hogy a közlésvégi irreguláris fonáció előfordulási gyakorisága a beszélő hangjának egy olyan jellemzője, amelyre emlékeznek a hallgatók) vizsgálatára. A hallgatók a próbák 64%-ában a beszélőre jellemző fonációs típust preferálták a beszélőre nem jellemző fonációs típussal szemben, ami szignifikánsan magasabb az 50% találgatási aránynál ($t(575)=7.122$; $p<0.0005$). Bár beszélőnként és hallgatónként elemezve némi változatosság figyelhető meg ennek a memória előhívásnak a hatékonyságában, az esetek jelentős részében a közlésvégi fonáció típusának szignifikáns hatása volt a válaszokra (mind a négy beszélő és a 9-ből 5 hallgató esetén).

Azokban az esetekben, amikor a pár transzformált tagjának alapfrekvenciája volt manipulált, a kísérleti személyek a próbák 86%-ában az eredeti felvételt preferálták. Ez konzisztens a szakirodalomban olvasható eredményekkel, amelyek azt mutatják, hogy az F0 átlag a beszélő személyének egy robusztus ismertetőjegye [28]. A zöngeminőség manipulálásának hatása és az F0 módosításának hatása között csak kis mértékű az interakció: a *zöngeminőség*+F_0 feltétel esetén a találati arány szignifikánsan magasabb volt, mint az F_0 feltétel esetén ($t(1150)=2.680$; $p=0.007$). Ez arra utal, hogy még egy olyan hatékony hangjellemző, mint az F0, jelenlétében is a megfelelő közlésvégi zöngeminőség könnyebbé teszi a hallgatók számára, hogy megállapítsák, melyik beszédminta felel meg a beszélő hangjának.

Következmény: *Ismerősök hangjának felismerésekor ez az információ a felismerést végző kognitív folyamat számára elérhető, ismertetőjegyként felhasználható.*

III.3. tézis: [C6] *Kimutattam, hogy a közlésvégi irreguláris fonáció előfordulási gyakoriságát ismeretlen beszélők esetén a kísérleti személyek már rövid perceptuális tanulás során eltárolják emlékezetükben.*



8. ábra: A beszélők szokásos zöngeminőségével, F0 átlagával, vagy mindkettővel rendelkező ingerre adott válaszok aránya a tanulás után. A vízszintes vonal az 50%-os találati aránynak felel meg, míg a függőleges szakaszok az átlaghoz tartozó konfidencia-intervallumokat jelölik.

Míg a III.2. tézis kísérletében a hallgatók jól ismerték a beszélők hangját, ebben a kísérletben olyan személyeket toboroztam (12 személy), akik nem ismerték a beszélőket. A kísérlet során azt vizsgáltam, hogy a beszélők hangjának rövid perceptuális tanulása során elsajátítható-e a rájuk jellemző közlésvégi irreguláris fonáció-gyakoriság.

A kísérlet egy tanulási szakaszból – amely során a kísérleti személyek egy videófelvételt láttak a négy beszélőről – és az azt követő három tesztszakaszból állt. Mindegyik tesztszakasz két tesztet tartalmazott. Az első azt mérte, hogy a hallgató személy képes-e felismerni a beszélőket a hangjuk alapján (ez lehetővé tette a hangok tanulásának nyomon követését). A második teszt nagyjából megegyezett a III.2. tézis kísérletében alkalmazott pár-összehasonlításos teszttel.

A beszélőfelismerési tesztek eredményei alapján az első tesztszakaszban a kísérleti személyek még nem ismerték eléggé a beszélők hangját, ezért az első pár-összehasonlításos tesztet gyakorlásnak tekintettem és adatait nem elemeztem. A másik két pár-összehasonlításos teszt eredményeit mutatja a 8. ábra, amelyek a III.2. tézis eredményeihez hasonló mintázatot mutatnak. A zöngeminőség feltételben a próbák 61%-ában a kísérleti személyek meg tudták állapítani, hogy a közlésvégi reguláris vagy irreguláris fonációval képzett beszédrészlet felel-e meg a beszélő hangjának (szignifikánsan magasabb, mint az 50% találgatási arány, $t = 5.328$; $p < 0.0005$). Ez csak akkor lehetséges, ha a hallgatók eltárolták emlékezetükben a beszélők jellegzetes fonációs szokásait.

Tehát az eredmények azt mutatták, hogy egy rövid, visszacsatolás nélküli perceptuális tanulás elegendő volt a hallgatóknak ahhoz, hogy a beszélők jellemző zöngeminőségét kódolják a beszélők hangjáról alkotott memóriaképükben.

Következmény (III.2. és III.3. tézis): *A beszélőmódosító technológiák és személyre szabott beszédszintézis rendszerek fejlesztésénél érdemes – a többi személyes akusztikus ismertetőjegy mellett – a beszélők fonációs szokásait is megfelelően transzformálni.*

5. Az eredmények alkalmazhatósága

A bevezetőben röviden ismertettem az irreguláris fonáció három lehetséges funkcióját a beszédkommunikációban (nyelvi információ és érzelmi állapot jelzése, valamint a beszélő személyére utalás). Az I. téziscsoportban ismertetett zöngeminőség osztályozó ennek megfelelően három területen alkalmazható. Egyrészt a beszédfelismerésben számos kiegészítő információt nyújthat: például hatékonyabbá teheti az intonációs frázisok határainak megtalálását, így lehetővé téve a bemenet kisebb egységekre darabolását (amelyeket hatékonyabban fel tud dolgozni a beszédfelismerő). Másodrészt a beszéd alapján a beszélő érzelmi állapotát osztályozó rendszerekben a zöngeminőség (reguláris vagy irreguláris) hasznos akusztikai jellemző lehet. Harmadrészt – építve a III.1. tézis eredményeire is – beszélőazonosító technológiákban a beszédjelből kinyerhető a közlésvégi irreguláris fonáció előfordulási gyakorisága, ami hozzájárulhat az egyes beszélők elkülönítéséhez.

A II. téziscsoport alkalmazási lehetőségeit is érdemes a zöngeminőség három feltételezett kommunikatív funkciója mentén sorba venni. Egyrészt szövegfelolvasó rendszerekben (text-to-speech, TTS) az előírt prozódiai struktúra megvalósítását természetesebbé teheti, ha a rendszer a megfelelő helyeken (például intonációs frázisok határán) irregulárissá alakítja a beszédet. Másodrészt, egyes érzelmi töltetekkel rendelkező beszéd előállításához valószínűleg szükséges a megfelelő helyeken a reguláris jel irregulárissá alakítása. Harmadrészt, a III.2. és a III.3. tézis eredményei alapján személyre szabott (egy adott beszélő hangjára adaptált) beszédszintetizátorok kialakításánál – egyéb egyéni jellemzők mellett – a beszélő fonációs szokásainak reprodukálása hozzájárulhat a meggyőző minőségű kimenet előállításához. A II.1 tézisben leírt eljárás hatékony alkalmazásához megterveztem és megvalósítottam egy szoftver eszközt. A grafikus felhasználói felület Nicolas Audibert (GIPSA-lab, Grenoble) munkája. A program az interneten ingyenesen elérhető [5].

A III. téziscsoport eredményei (fonetikai és pszicholingvisztikai jelentőségükön túl) rámutatnak, hogy az irreguláris fonáció kezelése (például az I. és II. téziscsoportban ismertetett eljárások segítségével) – mind magyar, mind angol nyelven – minőségjavuláshoz vezethet számos beszédtechnológiában.

[5] http://www.bohm.hu/glottalizer.html

A szerző saját publikációi

Folyóiratcikkek

[J1] Tamás Bőhm, Stefanie Shattuck-Hufnagel: Do listeners store in memory a speaker's habitual utterance-final phonation type? *Phonetica*, volume 66, issue 3, pp. 150-168, 2009.

[J2] Bőhm Tamás: Beszélőfelismerés – neurológiai háttér és pszichológiai modellek. *Magyar Pszichológiai Szemle* 62 (4), 2007. december, pp. 541-563.

[J3] Tamás Bőhm, Géza Németh: Algorithm for formant tracking, modification and synthesis, *Infocommunications Journal*, vol. LXII., 2007/1 Selected papers, pp. 15-20.

> [J3b] Bőhm Tamás, Németh Géza: Algoritmus formánsok követésére, módosítására és szintézisére, *Híradástechnika*, LXI. évfolyam, 2006/8, pp. 11-16.

[J4] Németh Géza, Olaszy Gábor, Bőhm Tamás, Ugron Zoltán: Szöveges adatbázis tervezése rendszerüzenet generátorhoz, *Híradástechnika*, LXI. évfolyam, 2006/3, pp. 38-42.

[J5] András Nagy, Péter Pesti, Géza Németh, Tamás Bőhm: Design issues of a corpus-based speech synthesizer. *Infocommunications Journal*, vol. LX., 2005/6 Selected papers, pp. 6-12.

> [J5b] Nagy András, Pesti Péter, Németh Géza, Bőhm Tamás: Korpusz-alapú beszédszintézis rendszerek megvalósítási kérdései. *Híradástechnika*, LX. évf., 2005/1, pp. 18-24.

[J6] Tamás Varga, Péter Benkő, Tamás Bőhm, Attila Eschwig-Hajts: Fluid simulation in telecommunication networks. *Infocommunications Journal*, vol. LVII, 2002/7 Selected papers, pp. 41-45.

Cikkek szerkesztett könyvekben

[B1] Bőhm Tamás, Ujváry István: Irreguláris fonáció előfordulása magyar beszédben, mint egyéni hangjellemző. In: Gósy Mária (szerk): *Beszédkutatás*, MTA Nyelvtudományi Intézet, Budapest, 2008. pp. 108-120.

[B2] Bőhm Tamás, Olaszy Gábor: A magyar [v] hang szerkezetének és zörejességének fonetikai vizsgálata. In: Gósy Mária (szerk): *Beszédkutatás*, MTA Nyelvtudományi Intézet, Budapest, 2007. pp. 19-34.

[B3] Bőhm Tamás: A glottalizáció szerepe a beszélő személy felismerésében. In: Gósy Mária (szerk): *Beszédkutatás*, MTA Nyelvtudományi Intézet, Budapest, 2006. pp. 197-207.

Konferenciacikkek

[C1] 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*, September 6-10, 2009, Brighton, United Kingdom, pp. 484-487.

[C2] Tamás Bőhm, Zoltán Both, Géza Németh: Automatic classification of regular vs. irregular phonation types, *Proc. NOLISP 2009*, June 25-27, 2009, Vic, Spain. pp. 53-61.

[C3] Tamás Bőhm, Nicolas Audibert, Stefanie Shattuck-Hufnagel, Géza Németh, Véronique Aubergé: Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles, *Proc. Acoustics 2008*, Paris. pp. 6141-6146.

[C4] Bőhm Tamás, Németh Géza: Egy egyszerű módszer modális beszéd glottalizálttá alakítására, *V. Magyar Számítógépes Nyelvészeti Konferencia*, 2007. december 6-7., Szeged, pp. 267-270.

[C5] Tamás Bőhm, Stefanie Shattuck-Hufnagel: Utterance-final glottalization as a cue for familiar speaker recognition, *Proc. Interspeech 2007*, August 27-31, 2007, Antwerp, Belgium, pp. 2657-2660.

[C6] Tamás Bőhm, Stefanie Shattuck-Hufnagel: Listeners recognize speakers' habitual utterance-final voice quality, *Proc. International Workshop on Paralinguistic Speech*, August 3, 2007, Saarbrücken, Germany, pp. 29-34.

[C7] Géza Németh, Géza Kiss, Tamás Bőhm: An algorithm for formant tracking, modification and re-synthesis of speech. *Trends in Speech Technology*, May 13-14, 2005, Cluj-Napoca, pp. 59-66.

Csak kivonatban megjelent konferencia-előadások

[C8] Tamás Bőhm: Is utterance-final glottalization a cue for speaker recognition by humans? *151st Meeting of the Acoustical Society of America*, June 5-9, 2006, Providence. Journal of the Acoustical Society of America, volume 119, issue 5, May 2006, pp. 3244-3245.

[C9] Tamás Bőhm, Géza Németh, Géza Kiss: A visual tool for the demonstration of formants in speech, *International Workshop in Phonetics Dedicated to the Memory of Farkas Kempelen*, March 11-13, 2004, Budapest.

Más hivatkozott források

[1] Fék, M., Pesti, P., Németh, G., Zainkó, Cs., "Generációváltás a beszédszintézisben," *Híradástechnika*, vol. LXI, no. 3, pp. 21-30, 2006.

[2] Tóth, B., Németh, G., "Rejtett Markov-modell alapú mesterséges beszédkeltés magyar nyelven," *Híradástechnika*, vol. LXIII, pp. 2-6, 2008.

[3] Szarvas, M., Fegyó, T., Mihajlik, P., Tatai, P., "Automatic Recognition of Hungarian: Theory and Practice," *International Journal of Speech Technology*, vol. 3, no. 3-4, pp. 237-251, 2000.

[4] K. Surana, "Classification of vocal fold vibration as regular or irregular in normal, voiced speech." M.Eng. thesis, MIT, 2006.

[5] Redi, L., Shattuck-Hufnagel, S., "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, no. 4, pp. 407-429, 2001.

[6] Henton, C., Bladon, A., "Creak as a sociophonetic marker," in *Language, speech and mind*. Hyman, L. M., Li, C. N., Eds. London: Routledge, pp. 3-29, 1987.

[7] Ishi, C. T., Sakakibara, K. I., Ishiguro, H., Hagita, N., "A Method for Automatic Detection of Vocal Fry," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 47-56, 2008.

[8] Gósy, M., *Fonetika, a beszéd tudománya*. Budapest: Osiris Kiadó, 2004.

[9] Fónagy, I., Magdics, K., *A magyar beszéd dallama*. Budapest: Akadémiai Kiadó, 1967.

[10] A. Markó, "A spontán beszéd néhány szupraszegmentális jellegzetessége." Ph.D. disszertáció, ELTE, 2005.

[11] Tóth, L., Kocsor, A., "A Magyar Telefonbeszéd-adatbázis (MTBA) kézi feldolgozásának tapasztalatai," in *Beszédkutatás 2003*. Gósy, M., Ed. Budapest: MTA Nyelvtudományi Intézet, pp. 134-146, 2003.

[12] Elekfi, L., "Hanglejtés," in *Nyelvművelő Kézikönyv,* Második kiad. Grétsy, L., Kovalovszky, M., Eds. Budapest: Akadémiai Kiadó, p. 774, 1983.

[13] Hirano, M., *Clinical examination of voice*. Vienna: Springer, 1981.

[14] Klatt, D. H., Klatt, L. C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820-857, Feb. 1990.

[15] Childers, D. G., Lee, C. K., "Vocal quality factors: analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394-2410, Nov. 1991.

[16] Bangayan, P., Long, C., Alwan, A. A., Kreiman, J., Gerratt, B. R., "Analysis by synthesis of pathological voices using the Klatt synthesizer," *Speech Communication*, vol. 22, no. 4, pp. 343-368, Sept. 1997.

[17] McCree, A. V., Barnwell, T. P., III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 4, pp. 242-250, 1995.

[18] Verma, A., Kumar, A., "Introducing Roughness in Individuality Transformation through Jitter Modeling and Modification," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05),* vol. 1, pp. 5-8, 2005.

[19] Loscos, A., Bonada, J., "Emulating rough and growl voice in the spectral domain," *7th International Conference on Digital Audio Effects (DAFx'04),* pp. 49-52, 2004.

[20] Olaszy, G., Bartalis, M., "Jelfeldolgozási és fonetikai algoritmusok kombinációja a gépi hanghatárjelölés javítására," in *Beszédkutatás 2008*. Gósy, M., Ed. Budapest: MTA Nyelvtudományi Intézet, pp. 208-220, 2008.

[21] Slifka, J., "Some physiological correlates to regular and irregular phonation at the end of an utterance," *Journal of Voice*, vol. 20, no. 2, pp. 171-186, June 2006.

[22] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L, and Zue, V., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, 1993.

[23] Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P. C., Goldman, S. L., "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *Journal of Speech and Hearing Research*, vol. 38, no. 6, pp. 1212-1223, Dec. 1995.

[24] Hanson, H. M., "Glottal characteristics of female speakers: Acoustic correlates," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466-481, 1997.

[25] Fawcett, T., "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[26] Kochanski, G., Grabe, E., Coleman, J., Rosner, B., "Loudness predicts prominence: Fundamental frequency lends little," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1038-1054, 2005.

[27] Moulines, E., Charpentier, F., "Pitch-Synchronous Wave-Form Processing Techniques for Text-To-Speech Synthesis Using Diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453-467, 1990.

[28] Abberton, E., Fourcin, A. J., "Intonation and speaker identification," *Language and Speech*, vol. 21, no. 4, pp. 305-318, Oct. 1978.