Az automatikus irreguláriszönge-detekció sikeressége az irregularitás mintázatának függvényében magyar (spontán és olvasott) beszédben

Markó Alexandra[1], Csapó Tamás Gábor[2]

[1] Eötvös Loránd Tudományegyetem, Fonetikai Tanszék
[2] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
marko.alexandra@btk.elte.hu, csapot@tmit.bme.hu

Kivonat: Az automatikus irreguláriszönge-detekció problémája előtérbe került az utóbbi évtizedekben. A jelen kutatásban a [10] algoritmust futtattuk le az irreguláris zönge előfordulásaira manuálisan felcímkézett, magyar nyelvű spontán és felolvasott beszédkorpuszokon, és azt vizsgáltuk, hogy 1. Mennyire pontos a pusztán akusztikai kulcsokon alapuló gépi detekció, és mennyire pontos az akusztikai és percepciós paramétereket egyaránt figyelembe vevő humán annotáció? 2. Milyen tényezők befolyásolják (rontják) az irreguláris zönge detekciójának sikerességét a gépi és a humán annotációkban? Eredményeink szerint az irregularitás általános célú annotációjára folyamatos szövegekben az automata algoritmus nagy hatásfokkal alkalmazható, mivel az előfordulások több mint 90%-át pontosan jelöli. A magánhangzók vizsgálatára létrehozott korpuszokban a gégezárhang figyelmen kívül hagyása miatt kevésbé pontos az automata detekció. Összességében az irregularitás alkalmazott definíciójától függ a gépi detektáló módszer hatásfoka.

1 Bevezetés

Az emberi beszédben reguláris (más néven 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 [5]. Az irreguláris fonáció a magyar nyelvben általában szakaszhatárokon (pl. mondat végén) [15] vagy magánhangzó–magánhangzó kapcsolatokban fordul elő [16]. Gyakran kíséri extrém alacsony alapfrekvencia és a glottális pulzusok gyors lecsökkenése [2]. Érzetileg recsegő, érdes jellegű beszédet jelent [3].

Az automatikus irreguláriszönge-detekció problémája előtérbe került az utóbbi évtizedekben, mivel mind a beszéddel kapcsolatos alapkutatásokban (fonetika), mind az alkalmazott kutatásokban (beszédtechnológia) szembesültek vele a kutatók, hogy az irreguláris zönge viszonylag gyakori jelenség. Egy 30 beszélős korpusz felolvasásaiban a szótagok 4,9–44,7%-a valósult meg részben vagy egészben irreguláris zöngével, a spontán beszédben az arányok 6,0 és 47,4% között szóródtak [18]. Az átlag 21,3%

volt az olvasott, 25,6% a spontán beszédben. A jelenség gyakoriságából adódóan a fonetikában leginkább az irreguláris zönge funkciói és más beszédparaméterekkel való összefüggésének kérdései állnak a kutatások középpontjában; míg a mesterséges beszéd-előállításban azért került az érdeklődés homlokterébe, mert a hangadatbázis minőségét több szempontból is befolyásolja.

Mivel a beszédtechnológia alapmódszereit idealizált beszédre dolgozták ki, az irreguláris zöngével képzett szakaszokon hibák léphetnek fel az automatikus F0-detekcióban, illetve spektrális elemzésben. Ugyanakkor a gépi szövegfelolvasó rendszerek minőségét javíthatja, ha a természetes beszédhez hasonlóan bizonyos pozíciókban (pl. szakaszhatárokon) modellezzük az irreguláris zöngét [8].

1.1 Automatikus irreguláriszönge-detekció

A zöngeminőség-osztályozók általában néhány, a beszédjelen mért akusztikai paraméter alapján hoznak döntést arról, hogy a zöngét reguláris vagy irreguláris zöngével képezték-e. Surana szupport vektor gép alapú osztályozást alkalmaz négy akusztikai jegyen [20]. Ishi és társai három másik jegy bevezetését javasolják, amelyek a beszédjel nagyon rövid szakaszában számolt teljesítményén alapulnak, és egyszerű küszöbértéket használnak a döntéshez [13]. Bőhm egyesíti az előző két osztályozót, és algoritmikus finomhangolással, valamint SVM alapú osztályozással javítja a pontosságot [5,6]. Kane és munkatársai 2013-ban publikálták szabadon elérhető automatikus irreguláriszönge-detektáló algoritmusukat [14], amely a beszéd lineáris predikció alapú maradékjelében méri a másodlagos csúcsok előfordulását és a kiugró, impulzusszerű csúcsokat. Az eljárást többféle irreguláris mintázaton, illetve több nyelven tesztelték, és bemutatták, hogy jobb eredmények érhetők el a korábbi irreguláriszönge-detektoroknál [10]. A fenti automatikus osztályozó eljárásokkal a reguláris és az irreguláris zöngével képzett beszéd közel tökéletesen elkülníthető egymástól.

1.2 A kutatás célja

A jelen kutatásban a [10] algoritmust futtattuk le az irreguláris zönge előfordulásaira manuálisan felcímkézett, magyar nyelvű spontán és felolvasott beszédkorpuszokon (interjúk, szövegfelolvasások, mondatolvasások, tipikus és diszfóniás beszélők). A kutatás célja annak megállapítása, hogy milyen mértékben illeszkedik egymáshoz a manuálisan és az automatikusan előállított címkesor. A kutatás kérdései: 1. Mennyire pontos a pusztán akusztikai kulcsokon alapuló gépi detekció, és mennyire pontos az akusztikai és percepciós paramétereket egyaránt figyelembe vevő humán annotáció? 2. Milyen tényezők befolyásolják (rontják) az irreguláris zönge detekciójának sikerességét a gépi és a humán annotációkban?

Hipotézisünk szerint a humán annotáció kevésbé pontos rövid időtartamú irreguláriszönge-előfordulások, valamint igen alacsony átlagos alaphangmagasság esetén, továbbá azokban az esetekben, amikor a zöngeminőség nemcsak az irregularitásban tér el a modálistól (pl. leheletes zönge). Az automatikus detektáló kapcsán

feltételezzük, hogy eredményességét befolyásolják a felvételi körülmények, a hangfelvétel minősége.

2 Módszerek

2.1 Hanganyag

A kutatásban többféle beszédkorpuszból válogattunk az irreguláris zöngére felcímkézett mintákat: felnőtt beszélők (egy 44 éves nő és egy 39 éves férfi) szövegfelolvasásait és interjúrészleteit a BEA adatbázisból [12]; diszfóniás (öt beszélő 21–38 év között) és kontroll (öt beszélő 20–24 év között) női beszélők mondatfelolvasásait [19]; valamint felnőtt beszélők (két – 25 és 31 éves – nő és két – 31 és 37 éves – férfi) mondatfelolvasásait a magánhangzó-kapcsolatok megvalósulásának elemzésére [17].

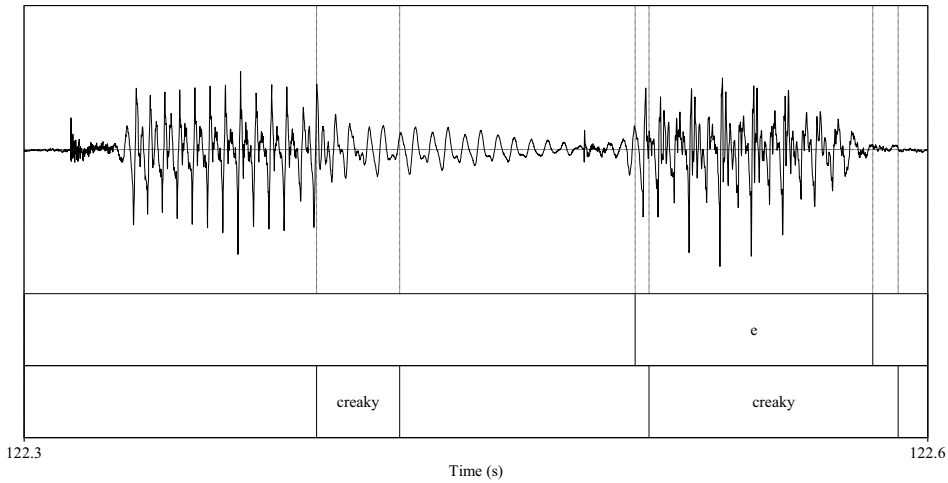
2.2 Irreguláriszönge-detekció

A beszédmintákon alkalmaztuk a CreakyDetection [10] irreguláriszönge-detektort, majd megvizsgáltuk, hogy a 10 ms-onként meghozott irreguláris/reguláris döntés milyen arányban felel meg a manuális irreguláriszönge-címkéknek.

2.3 Pontosság számítása

A célunk az volt, hogy egy-egy hangmintához meghatározzuk a manuális és automatikus címkézés pontosságát. Mivel azt feltételeztük, hogy mind a manuális, mind az automatikus címkézésben előfordulhat tévedés, ezért referenciának az összes irreguláris zönge címkével ellátott szakaszt vettük (azaz a manuális és automatikus címkék unióját). A manuális címkék határai tetszőlegesek lehetnek, míg az automatikus címkék 10 ms-os pontosságúak – emiatt az átfedő manuális/automatikus címkéket egyezőnek vettük. Az 1. ábrán egy példát mutatunk erre az esetre (az „e” és az alatta lévő „creaky” címke ugyanarra a magánhangzóra vonatkozik).

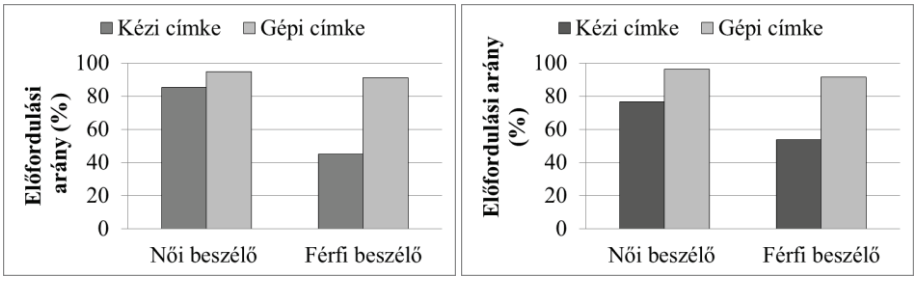
Hangmintánként kiszámítottuk a referenciához képest a manuális és automatikus címkék számát, majd ezt százalékos formába váltva kaptuk meg a pontosságot. Ez természetesen nem jelenti azt, hogy abszolút értelemben valóban a pontosságot határoztuk meg, hiszen az irregularitás meghatározása igen eltérő lehet (vö. pl. [1]).



1. ábra. Példa manuális (felső) és automatikus (alsó) irreguláris zönge címkére.

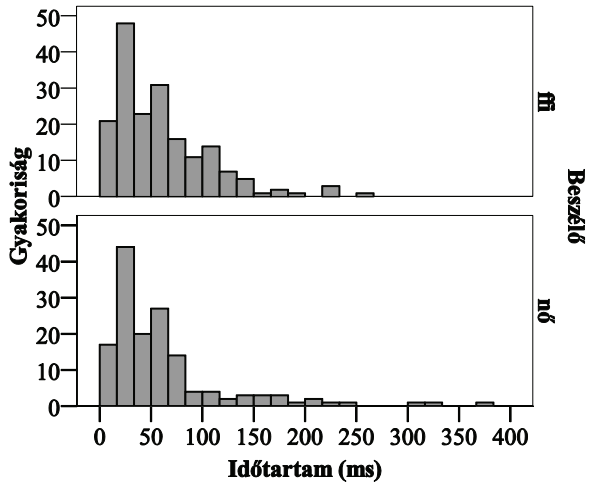
3 Eredmények

A BEA adatbázisból kiválasztott beszélők mindegyike valamilyen szempontból eltér az átlagostól, ezért beszédük címkézése az irreguláris zönge tekintetében nehezítettnek mondható. A női beszélő sokat glottalizálónak számít, felolvasásban a szótagjainak közel 30%-át, spontán beszédben 36%-át találtuk részben vagy egészben irregulárisnak. A férfi beszélő alaphangmagassága gyakran olyan mély tartományban valósul meg, hogy ez nehezíti a zöngeminőség manuális címkézését (az ő átlagos alaphangmagassága felolvasásban 100, spontán beszédben 88 Hz, de a 60–70 Hz közötti adatok sem ritkák a beszédében). A 2. ábrán látható, hogy ebben a két anyagban minden esetben az automatikus címkéző a pontosabb, kevés kivételtől eltekintve minden irreguláris szakaszt megtalált. A kézi címkézésnek a gépitől való eltérése több okra vezethető vissza (lásd alább), míg azoknak a manuálisan címkézett szakaszoknak a megjelenését, amelyeket az automata nem jelölt, általában az irregularitás értelmezésének különbségével magyarázhatjuk. Jelentős eltérés van a női és a férfi beszélő hanganyagának manuális címkézési eredményei között. Az előbbi, bár sokat glottalizál, az irreguláris szakaszok egyértelműen elkülöníthetők a beszédében, míg az utóbbi esetében gyakran nehéz megkülönböztetni az irreguláris zöngét az extrém alacsony alaphangmagasságtól.



2. ábra. A címkézés pontossága a BEA adatbázis két beszélője esetében: balra az olvasott, jobbra a spontán hanganyag eredményei.

Érdemes megvizsgálni azokat az eseteket, amelyeket a gépi annotáció jelölt, a kézi azonban nem. A manuális jelölésekre egyértelműen hatással van az irregularitás időtartama. Mintegy 50 ms-ban határozható meg a humán percepció küszöbértéke: a kézzel nem jelölt irreguláris szakaszok nagy része ennél rövidebb időtartamú (vö. 3. ábra). Ez természetesen nem jelenti azt, hogy a rövidebb irreguláris szakaszokat biztosan nem észleljük, csak azt, hogy ha valamilyen pozicionális marker nem teszi prominenssé, hajlamosak lehetünk elsiklani felettük.



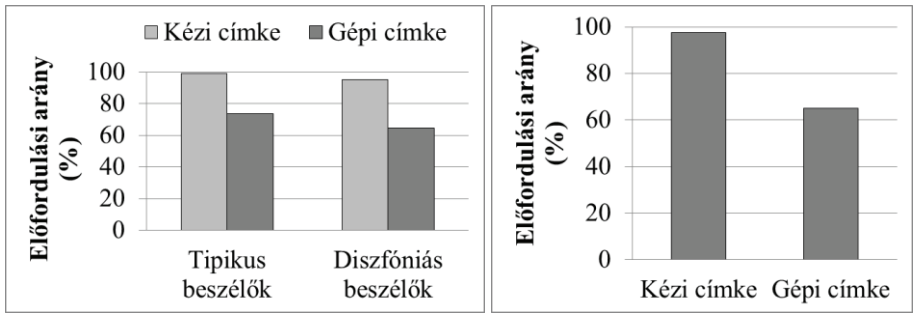
3. ábra. A kézi annotációból kimaradt gépi címkék gyakorisági eloszlása az irregularitás időtartama függvényében

Az eltérés tipikus oka még az, hogy az automatikus annotálás a zöngés obstruensek zárszakaszában, illetve a nem teljes zárfelpattanás esetén is irregularitást jelöl, akárcsak a pergőhang realizációiban. Ezek egy része ugyancsak rövid időtartamú címke. Egyes beszélőkre jellemző, hogy fojtott zöngét produkálnak, miközben nem artikulálnak – e felett a humán percepció elsiklik, az automata azonban ezekben is jelöli az

irregularitást. Találtunk néhány olyan esetet is, amikor az irregularitás látható ugyan a hullámformán, de a hallási percepció alapján nem észlelünk eltérést. Mivel [9] és [3] alapján akkor címkéztünk glottalizáltnak egy beszédrészletet, ha az akusztikai lenyomaton szemmel és auditív úton füllel egyaránt észlelhető volt az irregularitás, ezeket az eseteket a gépi annotációval való összevetés alapján sem tartanánk glottalizáltnak. Végül, különösen a hosszabb időtartamú címkék esetén, természetesen felmerül a humán címkéző figyelmetlensége, fáradása is.

Vizsgáltuk azon manuális címkék sajátosságait is, amelyekhez nem volt megfelelő gépi címke. Előfordult néhány esetben, hogy az utólagos ellenőrzés alapján ezeket nem tartanánk glottalizáltnak.

Egészen más eredményeket kaptunk a másik két korpuszból származó hanganyagok elemzésekor (4. ábra). Ezeknél nem minden egyes irreguláris szakasz címkézése volt a cél, hanem a magánhangzók, illetve magánhangzó-kapcsolatok irregularitásának címkézése. Tekintettel arra, hogy a felvételeken előre be volt jelölve ezeknek a szakaszoknak a pozíciója, a címkéző feladata az volt, hogy eldöntse, az adott magánhangzó(k)ban van-e irregularitás. Ennek alapján azt vártuk, hogy a manuális címkék aránya jobban meg fogja közelíteni a gépi címkékét, azonban ebben a két korpuszban a kézi címkék még felül is múlják számosságban az automatikus címkéket. Tekintettel arra, hogy a vizsgált esetek nem kis hányadában magánhangzós szókezdetek, illetve magánhangzó-kapcsolatok realizációi alkották az elemzés anyagát, a gégezárhang jelensége igen nagy arányban jelent meg az anyagban, ezeket azonban az automatikus címkézés egyetlen egy esetben sem jelölte, és az eljárást bemutató szakirodalmi forrás [10,14] sem utal arra, hogy ezeket az eseteket hogyan kezeli. Dilley és társai ugyanakkor egyértelműen az irregularitás altípusának veszik a gégezárhangot [9], sőt az újabb források szerint a gégezárhang és a glottalizáció valójában egyazon fiziológiai mechanizmussal jön létre [11]. A szerzők szerint gégezárhangot észlelünk, ha a csak egy periódusnyi tartamú, és glottalizációt hallunk, ha sorozatban több egymást követő periódusra kiterjed az ezeket létrehozó laringális konfiguráció.



4. ábra. A címkézés pontossága balra a patológiásbeszéd-korpuszban (magánhangzók jelölésében), jobbra a magánhangzó-kapcsolatok vizsgálatára létrehozott korpuszban.

Ugyanakkor ezekben a korpuszokban is megfigyelhető volt a humán percepció „túlműködése”, ugyanis néhány olyan esetben, amikor a hangzásbeli eltérés oka nem

az akusztikai jelben mérhető nagymértékű ingadozás, hanem például jellegzetes hangszínezet vagy leheletes zöngeképzés volt, a címkéző ezeket is jelölte.

4 Összefoglalás, következtetések

A vizsgálatunk arra irányult, hogy megállapítsuk, különböző magyar nyelvű korpuszokon milyen hatásfokkal alkalmazható a [10,14] által kifejlesztett automatikus irreguláriszönge-detektáló. Természetesen mindig az adott alkalmazástól és az irregularitás definíciójától függ, hogy a gépi elemző mennyire hatékony. Eredményeink alapján az irregularitás általános célú annotációjára folyamatos szövegekben az automata algoritmus nagy hatásfokkal alkalmazható, mivel az előfordulások több mint 90%-át pontosan jelöli. A humán percepció számára nehézséget okozó hangminőség (pl. alacsony alaphangmagasság) esetén is hatékony, és alkalmas az emberi tényező (figyelmetlenség, fáradás) ellensúlyozására. A gépi annotáció manuális ellenőrzésére ugyanakkor feltétlenül szükség van, mivel bizonyos beszédhangtípusok (pl. pergőhang, zöngés obstruensek) esetében szükségtelenül is jelöl, ugyanakkor a gégezárhangot – az algoritmus sajátosságaiból adódóan – nem ismeri fel. A tapasztalatok alapján az általános irreguláriszönge-címkézéshez érdemes 40-50 ms-os küszöbértéket beállítani. Így egyrészt a detektálónak az említett beszédhangtípusokra való érzékenysége is csökken, másrészt a címkék nagyobb mértékben korrelálnak a humán percepció által irregulárisnak minősített szakaszokkal.

A magánhangzók vizsgálatára létrehozott korpuszokban éppen a gégezárhang figyelmen kívül hagyása miatt kevésbé pontos az automata detektor, ugyanakkor ellenőrzésképpen való használata elkerülhetővé teszi, hogy a címkézést a humán percepciót zavaró tényezők (pl. leheletes zönge, feszített zöngeképzés) negatívan befolyásolják.

Köszönetnyilvánítás

A kutatást részben az SP2 Scopes Project on Speech Prosody támogatta.

Hivatkozások

1. Batliner, A., Burger, S., Johne, B., Kiessling, A.: MÜSLI: A classification scheme for laryngealizations. In: Working Papers, Prosody Workshop, Lund, Schweden (1993) 176–179
2. Blomgren, M., Chen, Y., Ng, M. L., Gilbert, H. R.: Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. Journal of the Acoustical Society of America 103 (1998) 2649–2658
3. Bőhm, T., Ujváry, I.: Az irreguláris fonáció mint egyéni hangjellemző a magyar beszédben. Beszédkutatás 2008 (2008) 108–120

4. Bőhm, T., Audibert, N., Shattuck-Hufnagel, S., Németh, G., Aubergé, V.: Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. In: Acoustics'08 (2008) 6141–6146
5. Bőhm, T.: Analysis and modeling of speech produced with irregular phonation. PhD disszertáció, BME TMIT (2009)
6. Bőhm, T., Both, Z., Németh, G.: Automatic Classification of Regular vs. Irregular Phonation Types. In: NOLISP, (2009) 43–50
7. Collins, B., Mees, I. M.: Practical phonetics and phonology: A resource book for students. Routledge, New York (2008)
8. Csapó, T. G., Németh, G.: Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation. IEEE Journal on Selected Topics in Signal Processing 8 (2014) 209-220
9. Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M.: Glottalization of word-initial vowels as a function of prosodic structure. Journal of Phonetics 24 (1996) 423–444
10. Drugman, T., Kane, J., Gobl, C.: Data-driven Detection and Analysis of the Patterns of Creaky Voice. Computer Speech and Language 28 (2014) 1233–1253
11. Esling, J. H., Harris, J. G.: States of the glottis: an articulatory phonetic model based on laryngoscopic observations. In: Hardcastle, W. J. – Mackenzie Beck, J., eds.: A figure of speech: A festschrift for John Laver. Lawrence Erlbaum Association, Mahwah (2005) 347–383
12. Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T. E., Beke, A., Neuberger, T., Nikléczy, P.: BEA: beszélt nyelvi adatbázis. In: Gósy, M., ed.: Beszéd, adatbázis, kutatások. Budapest, Akadémiai Kiadó (2012) 9–24
13. Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., Hagita, N.: A Method for Automatic Detection of Vocal Fry. IEEE Transactions on Audio, Speech and Language Processing 16 (2008) 47–56
14. Kane, J., Drugman, T., Gobl, C.: Improved automatic detection of creak. Computer Speech and Language 27 (2013) 1028–1047
15. Markó, A.: A glottalizáció határjelző szerepe a felolvasásban. Beszédkutatás 2011 (2011) 31–45
16. Markó, A.: Az irreguláris zönge szerepe a magánhangzók határának jelölésében V(#)V kapcsolatokban. Beszédkutatás 2012 (2012) 5–29
17. Markó, A.: Boundary marking in Hungarian V(#)V clusters with special regard to the role of irregular phonation. The Phonetician 103 (2012) 7–26
18. Markó, A.: Az irreguláris zönge funkciói a magyar beszédben. Budapest: ELTE Eötvös Kiadó, (2013)
19. Markó, A.: Glottalizáció és diszfónia. Gyógypedagógiai Szemle XLII/1 (2014) 23–36
20. Surana, K.: Classification of vocal fold vibration as regular or irregular in normal voiced speech. MS thesis, MIT, USA, (2006)