Folytonos paraméterű vokóder rejtett Markov-modell alapú beszédszintézisben – magyar nyelvű kísérletek 12 beszélővel

Csapó Tamás Gábor, Németh Géza

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék,
e-mail: {csapot,nemeth}@tmit.bme.hu

Kivonat A jelen cikkben egy vokódert mutatunk be, amely a beszédet folytonos paraméterekként reprezentálja. A módszer a szakirodalomban ismert parametrikus vokóderekhez képest két fő tulajdonságban különbözik: 1) a zöngés és zöngétlen szakaszokat egységesen (a gerjesztőjel explicit megkülönböztetése nélkül) kezeljük időtartományban egy folytonos alapfrekvencia-mérő algoritmus használatával, 2) a gerjesztőjelet frekvenciatartományban zöngés és zöngétlen komponensek összegeként állítjuk elő, melyeket egy maximális zöngésségi frekvencia érték határol. A vokóder így csak folytonos paramétereket alkalmaz, ami a statisztikai modellezés szempontjából kedvező. Mivel a szintézis rész számításigénye alacsony, ezért a javasolt vokóder hatékonyan alkalmazható korlátozott erőforrású eszközökön is (pl. Android okostelefon) rejtett Markov-modell alapú beszédszintézisben. Az új vokódert beszélő adaptációban is teszteltük, mellyel tetszőleges beszélőre emlékeztető beszédszintetizátor hangot tudunk létrehozni.

Kulcsszavak: gépi tanulás, beszédtechnológia, statisztikai modellezés

1. Bevezetés

A gépi szövegfelolvasás (TTS, Text-To-Speech) egyik legkorszerűbb technológiája a statisztikai parametrikus beszédszintézis [1]. A beszédtechnológiában a statisztikai parametrikus módszerekhez gyakran alkalmazzák a rejtett Markov-modelleket (HMM) [2,3]. Zen és társai szerint három fő területen van kutatásra szükség ahhoz, hogy a statisztikai parametrikus TTS módszerek a természeteshez közeli beszédet eredményezzenek: 1) új típusú vokóderek, 2) az akusztikai modellek pontossága, 3) és a paraméterek túlsimítottsága [1]. Jelen cikkben az első területtel foglalkozunk.

1.1. Vokóderek a statisztikai parametrikus beszédszintézisben

A szakirodalomban számos beszédkódoló módszerről olvashatunk, melyeknek eredeti célja a beszéd paraméterekre bontása (kódolás, analízis lépés) azért, hogy

a távközlési csatornán minél kisebb sávszélesség mellett lehessen átvinni a jelet (beszédet) [4, 244. o.]. Az átvitel után, a vevő oldalon a paramétereket visszaalakítják beszédjellé (dekódolás, szintézis lépés). A parametrikus kódolók, azaz vokóderek családjába tartozik az LPC (Linear Predictive Coding) kódoló, valamint ennek továbbfejlesztett változatai, melyek az elsődleges cél mellett alkalmasak a beszédjel tulajdonságainak változtatására is (pl. F0 módosítás).

Az elmúlt évtizedekben számos vokóder típust kidolgoztak, melyeket a következő kategóriákba sorolhatunk: kevert gerjesztés [5], glottális forrás alapú módszerek [6,7,8], harmonikus-zaj alapú módszerek [9] és maradékjel alapú módszerek [10,11,12] (teljes összehasonlítás: [13, Introduction]). Mindegyik fenti vokódernek az a célja, hogy a HMM-TTS korai változataiban alkalmazott impulzus-zaj elvű vokóder robotosságát, gépiességét, 'zizegését' csökkentsék. Ugyan léteznek olyan vokóderek, melyek közel természetes beszédet tudnak visszaállítani, de ezek tipikusan magas számításigényűek, és ezért nem alkalmazhatóak valós időben (pl. STRAIGHT, [14]).

1.2. A jelen kutatás

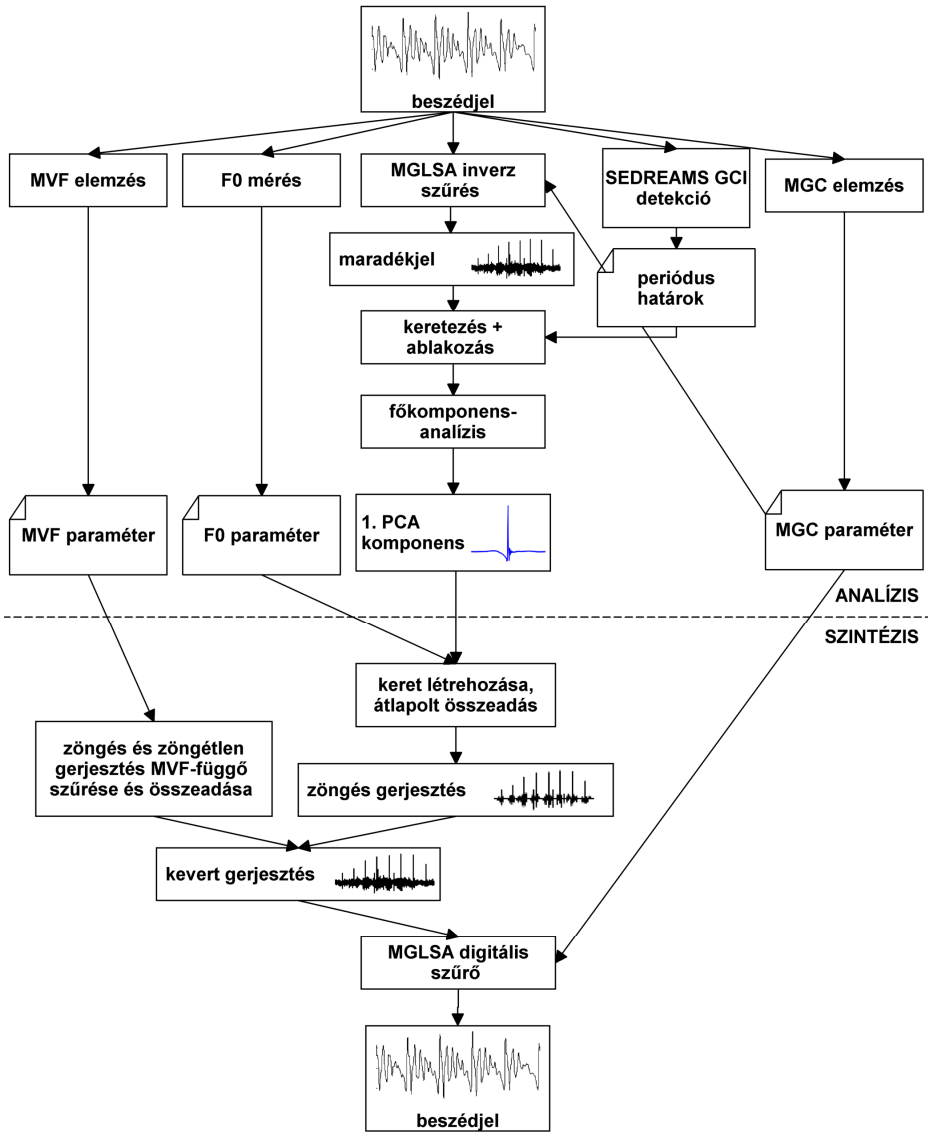
A jelen cikkben egy alacsony komplexitású és számításigényű vokódert mutatunk be. A vokóder korábbi változata maradékjel alapú, és folytonos alapfrekvenciát valamint maximális zöngésségi frekvenciát alkalmaz a zöngés és zöngétlen beszédhangok egységes modellezésére [15]. Később ezt tovább javítottuk a zöngétlen hangok frekvenciakomponenseinek optimális súlyozásával [16]. A mostani cikkben csak a vokóder legutolsó változatát ismertetjük [13] és az erre épülő beszédszintézis alkalmazásokat (magyar nyelvű beszédszintézis Android okostelefonon; TTS adott beszélőre adaptálása néhány percnyi hangminta alapján) is bemutatjuk.

2. Folytonos paraméterű vokóder

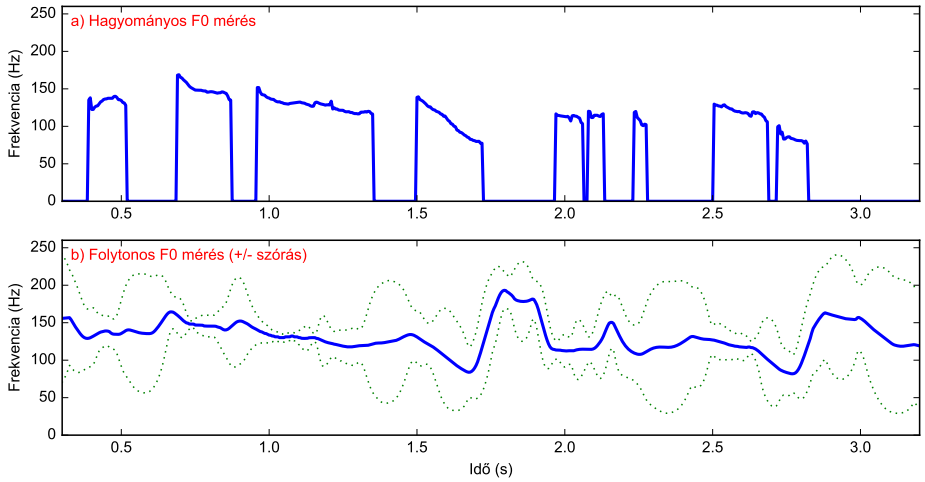
A vokóder analízis és szintézis részekből áll. Az analízis lépés a beszédjel alapján gerjesztési- és spektrális paramétereket állít elő, melyeket a rejtett Markov-modell alapú beszédszintézis modelljeinek betanításához lehet felhasználni. A HMM modell eredményeképpen tetszőleges bemeneti szöveghez generálni tudjuk a gerjesztési- és spektrális paramétereket, majd a vokóder szintézis lépésében a beszéd visszaállítható ezekből.

2.1. Analízis

Az analízis lépéseit az 1. ábra szaggatott vonal feletti része mutatja. Az analízis rész bemenete beszéd hullámforma, amelyet 7,6 kHz-es aluláteresztő szűrés után 16 kHz mintavételezéssel és 16 bites lineáris PCM kvantálással tárolunk. A beszédjelen egy folytonos alapfrekvencia detektorral [17,18] 5 ms eltolással kiszámítjuk az F0 paramétert (F0cont). Ez az F0 detektor a zöngétlen szakaszokon interpolálja az F0-t és Kálmán-szűrést alkalmaz, melynek eredményére a



1. ábra. Analízis (a szaggatott vonal felett) és szintézis (a szaggatott vonal alatt) a folytonos paraméterű vokóderrel.



2. ábra. Az F0 mérés eredménye a) a Snack hagyományos F0 számító algoritmussal [21], b) az SSP folytonos F0 számító algoritmussal [17,18]. A kék folytonos vonal az F0 kontúr, míg a zöld pontozott vonalak a +/- szórást jelölik.

2. ábra mutat példát. Ezután a 'maximális zöngésségi frekvencia' (Maximum Voiced Frequency, MVF, [19]) számítása következik. A következő lépésben spektrális elemzést végzünk 'mel-általánosított kepsztrum' (Mel-Generalized Cepstrum, MGC, [20]) módszerrel . Az elemzéshez 24-ed rendű MGC-t számítunk $\alpha = 0{,}42$ és $\gamma = -1/3$ értékekkel. Végül az MGLSA inverz szűréssel kapott maradékjel zöngeszinkron periódusaiból főkomponens-analízisével kinyerünk egy a későbbi szintézishez használható gerjesztőjelet ('PCA maradékjel', részletek: [15]).

2.2. Az új vokóder rejtett Markov-modell alapú beszédszintézisben

Az analízis résznél leírt paramétereket (F0cont, MVF és MGC) kiszámítjuk a tanító beszédadatbázis mondatainak minden keretére, 5 ms-os eltolással. Az F0cont és MVF paramétereket logaritmizáljuk, majd az MGC-vel együtt a derivált és második derivált értékeket is eltároljuk a paraméterfolyamban. Mivel a paraméterek folytonosak (azaz nincs bennük szakadás, mint a hagyományos F0 kontúr esetén), a modellezés hagyományos HMM-ekkel történik. A tanítás többi része (pl. környezetfüggő címkézés, döntési fák, időtartamok modellezése) a HTS-HUN rendszerrel megegyező módon történik [2,22].

2.3. Szintézis

A szintézis lépéseit az 1. ábra szaggatott vonal alatti része mutatja be. A szintézis bemenetei az analízis eredménye után gépi tanulással modellezett paraméterek (F0cont, MVF és MGC) illetve a 'PCA maradékjel'. A visszaállítás során először a 'PCA maradékjelet' átlapoltan összeadjuk az F0cont-tól függő

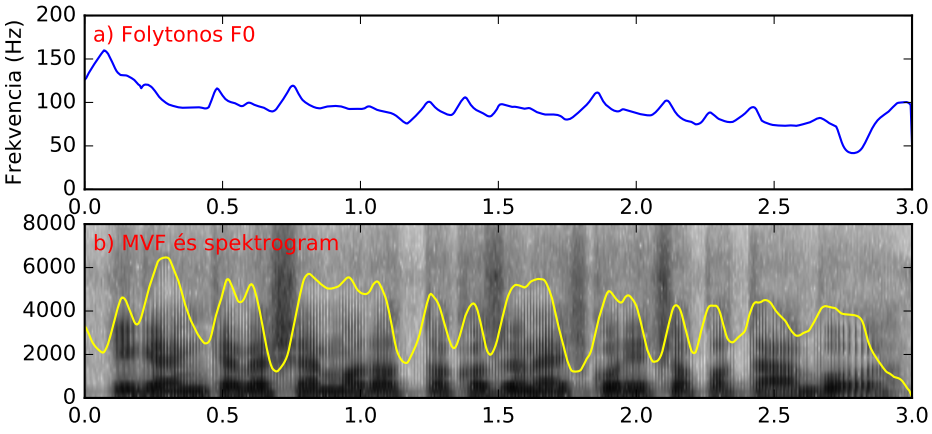
távolságra, ami a gerjesztés zöngés komponensét adja meg. A zöngétlen komponenst fehérzajból hozzuk létre. Mivel nincs külön zöngés/zöngétlen paraméter folyam, az MVF paraméter modellezi a zöngésségi információt, melyet a 3. ábra mutat: a zöngétlen beszédhangok esetén az MVF általában alacsony (200–500 Hz körüli), a zöngés beszédhangoknál magas (tipikusan 4 kHz fölötti), míg a kevert gerjesztésű beszédhangoknál a két véglet közötti (pl. zöngés réshangok esetén 2–3 kHz közötti). A zöngés gerjesztést keretenként az MVF-től függő aluláteresztő szűrővel, míg a zöngétlen gerjesztést felüláteresztő szűrővel módosítjuk, majd összeadjuk a két gerjesztési komponenst. Végül a a szintetizált beszédet az összeadott kevert gerjesztés alapján előállítjuk MGLSA szűréssel az MGC paramétereket felhasználva [23]. Az így szintetizált beszédre a 3. ábra mutat egy példát.

3. Kísérletek és eredmények

3.1. Beszélőfüggő tanítás 12 beszélővel és átlaghang

A kísérletek során magyar nyelvű mintákon végeztük a HMM-ek tanítását és minta szövegek szintézisét. Ehhez a nyelvspecifikus lépéseket a HTS-HUN rendszerből kiindulva alkalmaztuk [22]. A PPBA adatbázis [24,25] hat férfi és hat női beszélőjének hanganyagával végeztünk beszédszintézis kísérleteket. Ehhez a teljes, kb. 2 órányi (beszélőnként közel 2000 mondat) beszédfelvételt és a hozzá tartozó címkézést használtuk fel beszélőfüggő tanítás keretében.

A beszélőfüggő kísérletek után átlaghangot [26] is készítettünk az új vokóderrel és a HTS-HUN rendszerrel. Ehhez a PPBA adatbázis 10 beszélőjét használtuk fel három különböző módon: 1) a tíz beszélőtől származó átlaghang, 2) öt férfi beszélőtől származó átlaghang, 3) öt női beszélőtől származó átlaghang.



3. ábra. Szintetizált beszédminta egy férfi beszélőtől: *'Igen kevesen maradtak az Ön egykori csapatából.'*

1. táblázat. A meghallgatásos teszt eredménye.

Férfi beszélők								Női beszélők							
	1.	2.	3.	4.	5.	6.	7.		1.	2.	3.	4.	5.	6.	7.
FF1	0	0	1	1	0	2	1	NŐ1	0	0	0	0	1	1	3
FF2	1	1	0	3	0	0	0	NŐ2	0	1	1	1	0	2	0
FF3	**4**	1	0	0	0	0	0	NŐ3	0	**3**	0	1	1	0	0
FF4	0	0	1	0	1	1	2	NŐ4	0	1	0	0	3	1	0
FF5	0	2	0	1	2	0	0	NŐ5	**5**	0	0	0	0	0	0
FF6	0	0	0	0	1	2	2	NŐ6	0	0	1	1	0	1	2
FF_átlag	0	1	3	0	1	0	0	NŐ_átlag	0	0	3	2	0	0	0

3.2. Meghallgatásos teszt

A 12 beszélőtől valamint a férfi és női átlaghangból 100–100 mondatot szintetizáltunk, majd egy bekezdést kiválasztottunk egy internetes meghallgatásos teszthez. A tesztelők feladata az volt, hogy ugyanazon mondatokat meghallgatva az összes beszélőtől eldöntsék, hogy melyik férfi és melyik női bemondót preferálják (azaz sorba kellett állítani a beszélőket aszerint, hogy melyik hangkarakter tetszett a legjobban). A preferenciatesztben 5 beszédtechnológiai szakértő vett részt (30–70 év közötti férfiak). Az eredményeket az 1. táblázat mutatja, mely szerint a nők közül NŐ5 és NŐ3 az előnyben részesített, míg a férfiak közül FF3. Az előbbinek az lehet az oka, hogy a preferált női beszélők professzionális bemondók, így az ő hangjuk várhatóan előnyösebb éles TTS rendszerben.

3.3. Beszélő adaptáció

Készítettünk egy Android okostelefonos alkalmazást, amely új beszélőktől hangminták gyűjtésére alkalmas. Öt beszélőtől gyűjtöttünk ilyen módon okostelefonon / tableten felolvasott hangmintákat (50–50 mondatot), majd beszélő adaptációt [22,26] indítottunk az átlaghangokat felhasználva (3.1. fejezet). Az informális meghallgatások szerint az 5 beszélős átlaghangokkal adaptált minták jobban emlékeztetnek az eredeti beszélőre, mint a 10 beszélős átlaghanggal adaptáltak, valószínűleg azért, mert a külön férfi és női beszélőkből álló átlaghangok jobban megőrzik az adott nem jellemzőit.

3.4. Androidos implementáció

Az új vokódert a HTS-HUN rendszer alacsony erőforrású eszközökre optimalizált változatához illesztettük [27]. A HMM-TTS az új vokóderrel közel valós időben (néhány 10 ms-on belül) képes szövegből beszédet szintetizálni átlagos Androidos telefonokon. Precíz meghallgatásos tesztet nem végeztünk az okostelefonokon, de a tapasztalatok szerint az új, folytonos paraméterű vokóderrel kellemesebb beszéd szintetizálható, mint a HTS rendszer egyszerű impulzus-zaj gerjesztésű vokóderével. Korábbi internetes percepciós tesztekben már igazoltuk, hogy az új vokóder természetesebb, mint az alaprendszer [13,15,16].

4. Következtetések

Kutatásunk eredményei számos beszédtechnológiai alkalmazásban felhasználhatóak, amelyek egyrészt hozzájárulhatnak a természetesebb ember-gép kommunikációhoz, másrészt segíthetnek megérteni az emberi beszédképzés működését. A bemutatott beszédszintetizátor rendszer javítja a korlátozott erőforrású eszközökben (pl. Android okostelefon) alkalmazott gépi szövegfelolvasás minőségét. A kevés erőforrás miatt bonyolultabb gerjesztési modellek nehézkesen kezelhetőek, viszont a legújabb vokóder a korlátozott erőforrású eszközökön is képes közel valós idejű beszédszintézisre. A beszédsérülteket segítő kommunikációs eszközökben hasznos lehet, ha a rendszer az eredeti beszélőre emlékeztető hangon szólal meg, amit a beszélő adaptációval oldhatunk meg.

Köszönetnyilvánítás

A kutatást részben támogatta a SCOPES projekt (SP2: SCOPES project on speech prosody, SNSF no IZ73Z0_152495-1) és a VUK (AAL-2014-1-183) projekt keretében az Európai Unió és a Nemzeti Kutatási, Fejlesztési és Innovációs Alap.

Hivatkozások

1. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication **51**(11) (2009) 1039–1064
2. Tóth, B.P.: Rejtett Markov-modell alapú gépi beszédkeltés. PhD disszertáció, BME TMIT (2013)
3. Tóth, B.P., Németh, G.: Rejtett Markov-modell alapú szövegfelolvasó adaptációja félig spontán magyar beszéddel. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2009), Szeged, Magyarország (2009) 246–256
4. Németh, G., Olaszy, G., eds.: A MAGYAR BESZÉD; Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek. Akadémiai Kiadó, Budapest (2010)
5. Yoshimura, T., Tokuda, K.: Mixed excitation for HMM-based speech synthesis. In: Proc. Eurospeech, Aalborg, Denmark (2001) 2263–2266
6. Cabral, J.P., Renals, S., Yamagishi, J., Richmond, K.: HMM-based speech synthesiser using the LF-model of the glottal source. In: Proc. ICASSP, Prague, Czech Republic (2011) 4704–4707
7. Degottex, G., Lanchantin, P., Roebel, A., Rodet, X.: Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. Speech Communication **55**(2) (2013) 278–294
8. Raitio, T., Suni, A., Vainio, M., Alku, P.: Comparing glottal-flow-excited statistical parametric speech synthesis methods. In: Proc. ICASSP, Vancouver, Canada (2013) 7830–7834
9. Erro, D., Sainz, I., Navas, E., Hernáez, I.: Improved HNM-based Vocoder for Statistical Synthesizers. In: Proc. Interspeech, Florence, Italy (2011) 1809–1812
10. Drugman, T., Dutoit, T.: The Deterministic Plus Stochastic Model of the Residual Signal and its Applications. IEEE Transactions on Audio, Speech and Language Processing **20**(3) (2012) 968–981

11. Drugman, T., Raitio, T.: Excitation Modeling for HMM-based Speech Synthesis: Breaking Down the Impact of Periodic and Aperiodic Components. In: Proc. ICASSP, Florence, Italy (2014) 260–264
12. Wen, Z., Tao, J.: Amplitude spectrum based Excitation model for HMM-based Speech Synthesis. In: Proc. Interspeech, Portland, Oregon, USA (2012) 1428–1431
13. Csapó, T.G., Németh, G., Cernak, M., Garner, P.N.: Parametric Vocoder with Continuous F0 Modeling and Residual-based Excitation for Speech Synthesis. submitted to Speech Communication (2017)
14. Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication **27**(3) (1999) 187–207
15. Csapó, T.G., Németh, G., Cernak, M.: Residual-Based Excitation with Continuous F0 Modeling in HMM-Based Speech Synthesis. In Dediu, A.H., Martín-Vide, C., Vicsi, K., eds.: Lecture Notes in Artificial Intelligence. Volume 9449. Springer International Publishing, Budapest, Hungary (2015) 27–38
16. Csapó, T.G., Németh, G., Cernak, M., Garner, P.N.: Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder. In: Proc. EUSIPCO, Budapest, Hungary (2016) 1338–1342
17. : Speech Signal Processing - a small collection of routines in Python to do signal processing [Computer program] (2015) https://github.com/idiap/ssp.
18. Garner, P.N., Cernak, M., Motlicek, P.: A simple continuous pitch estimation algorithm. IEEE Signal Processing Letters **20**(1) (2013) 102–105
19. Drugman, T., Stylianou, Y.: Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra. IEEE Signal Processing Letters **21**(10) (2014) 1230–1234
20. Tokuda, K., Kobayashi, T., Masuko, T., Imai, S.: Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. In: Proc. ICSLP, Yokohama, Japan (1994) 1043–1046
21. Talkin, D.: A Robust Algorithm for Pitch Tracking (RAPT). In Kleijn, W.B., Paliwal, K.K., eds.: Speech Coding and Synthesis. Elsevier (1995) 495–518
22. Tóth, B.P., Németh, G.: Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis. Acta Cybernetica **19**(4) (2010) 715–731
23. Imai, S., Sumita, K., Furuichi, C.: Mel Log Spectrum Approximation (MLSA) filter for speech synthesis. Electronics and Communications in Japan (Part I: Communications) **66**(2) (1983) 10–18
24. Olaszy, G.: Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai. Beszédkutatás 2013 (2013) 261–270
25. Tóth, B.P., Németh, G., Olaszy, G.: Beszédkorpusz tervezése magyar nyelvű, rejtett Markov-modell alapú szövegfelolvasóhoz. Beszédkutatás 2012 **20** (2012) 278–295
26. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm. IEEE Transactions on Audio, Speech, and Language Processing **17**(1) (2009) 66–83
27. Tóth, B.P., Németh, G.: Optimizing HMM Speech Synthesis for Low-Resource Devices. Journal of Advanced Computational Intelligence and Intelligent Informatics **16**(2) (2012) 327–334