Ahots-sintesirako HNM ereduan oinarritutako vocoder hobetua

D. Erro, I. Sainz, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, I. Odriozola

AHOLAB Signal Processing Laboratory, Euskal Herriko Unibertsitatea, Bilbao
derro@aholab.ehu.es, inaki@aholab.ehu.es, eva@aholab.ehu.es, inma@aholab.ehu.es, ion@aholab.ehu.es, ibon@aholab.ehu.es, igor@aholab.ehu.es

Abstract

Statistical parametric synthesizers have achieved very good performance scores during the last years. Nevertheless, as they require the use of vocoders to parameterize speech (during training) and to reconstruct waveforms (during synthesis), the speech generated from statistical models lacks some degree of naturalness. In previous works we explored the usefulness of the harmonics plus noise model in the design of a high-quality speech vocoder. Quite promising results were achieved when this vocoder was integrated into a synthesizer. In this paper, we describe some recent improvements related to the excitation parameters, particularly the so called maximum voiced frequency. Its estimation and explicit modeling leads to an even better synthesis performance as confirmed by subjective comparisons with other well-known methods.

Laburpena

Azken urteotan ahots-sintetizagailu estatistiko eta parametrikoek oso emaitza onak lortu dituzte. Hala ere, ahotsa parametrizatzeko (entrenamendu-fasean) eta berreraikitzeko (sintesi-fasean) vocoderrak erabili behar direnez, eredu estatistikoetatik abiatuta sortzen den ahotsak naturaltasun falta erakusten du. Aurreko lan batzuetan harmonikoak gehi zarata ereduaren erabilgarritasuna aztertu genuen kalitate handiko vocoderra diseinatzeko. Vocoder hau ahotsa sortzeko sistema batean integratu zenean etorkizun handiko emaitzak lortu ziren. Artikulu honetan kitzikadura parametroetan egindako hobekuntza, batez ere ahostun frekuentzia handienaren erabilera deskribatzen da. Frekuentzia honen estimazioak eta modelatzeak sintesi-emaitzak hobetzen dituzte, beste metodo oso ezagun batzuekin konparatzean egiaztatu dena.

Keywords: Vocoder, statistical parametric speech synthesis, harmonics plus noise model, speech parameterization

Hitz gakoak: Vocoder, ahots-sintesi parametriko eta estatistikoa, harmonikoak eta zarata eredua, ahots-parametrizazioa

1. Sarrera

Azken urteotan sintesi estatistiko eta parametrikoa (Zen et al., 2009) nagusitu da. Sintesi-sistema estatistikoek fonemen ezaugarri akustikoak eta iraupena ingurune menpeko Markov ezkutuko ereduen bidez modelatzen dituzte (Context Dependent Hidden Markov models, CDHMM). Sintesia egiteko momentuan eta sortu behar den esaldiaren deskriptore fonetiko eta linguistikoak izanda, dagozkion fonema mailako CDHMMak kateatzen esaldi mailako HMM sortzen dute. Orduan ahots-seinalea berreraikitzen dute ~~esaldi~~ HMMrekiko egiantz handieneko esaldia daukan bektore akustiko segidatik abiatuta. Esparru estatistiko honen abantailen artean, honako hauek aipatzea merezi du: 1) sortzen diren ahots-sintesirako sistemek oinatz txikia daukate eta hau oso komenigarria da dispositibo txikietan erabiltzeko. 2) sistema hauek oso malguak dira: ahotsa ereduak erabiliz sortzen denez, sistema mota honek eredu hauek aldatzeko gai den edozein teknikari etekina atera diezaioke (egokitzapena, interpolazioa etab.). Hori horrela izanda, ahotsaren ezaugarri akustikoak, hitz egiteko modua edo emozioa erraz alda daitezke. 3) sistema hizkuntza berrietarako egokitzea nahiko erraza da. 4) sintesi-sistema estatistikoek sortzen duten ahotsa leunagoa da.

Artikulu honetan sintesi parametrikoak daukan erronkarik garrantzitsuenetariko bat tratatzen da, hau da, kalitate handiko vocoderraren diseinua. Lan inguru honetan vocoderrak erabiltzen dira bai entrenamendu korpuseko ahots-seinaleak ereduan eraikitzeko erabiltzen diren bektoreak bihurtzeko eta bai sistemak sortzen dituen parametro-bektoreen segidak erabiliz ahotsa berreraikitzeko. Horregatik sintesi-sistemaren performantzia oso lotuta dago vocoderraren kalitate mailarekin. Arrazoi honengatik batez ere, entzuleek nahiago dituzte unitate hautaketa erabiltzen duten sistema onen seinaleak, sintesi estatistiko sistema onenak baino. (Mendez et al. 2010).

Orokorrean ahots seinaleak parametrizatzen dira bizpahiru bektore korronte erabiliz: bat f0-rako, bat espektrorako eta hautazko bat kitzikapenerako. Espektroari dagokionez, MFCC (*Mel Frequency Cepstral Coefficients*) eta LSP (*Line Spectral Pairs*) erabiltzen dira, gehien. Koefiziente hauek kalkulatzeko erabiltzen den metodoa sistemaren arabera aldatzen da, baina gehienek Straight (Kawahara, 2006) edo Mel-hedatu analisi cepstrala (Tokuda et al., 1994) erabiltzen dute. Kitzikapenaren parametrizazioari dagokionez, azken urteotan proposamen ugari izan da: kitzikapen mistoa ahostun indar-guneak (Yoshimura et al., 2001) (Gonzalvo et al., 2007) edo ez-periodikotasunaren neurketak eta fase-aldaketak (Zen et al., 2007) kontuan hartzen duena, pultsu eta zaratarako egoeraren menpeko iragazkiak (Maia et al., 2007), hondakina eredu deterministaren eta estokastikoaren bidez modelatzea

(Drugman et al., 2009), iturri glotalaren modelatzea (Raitio et al., 2011), etab.

Harmoniko gehi zarata (*Harmonics plus Noise Model*, HNM) ereduan oinarrituta dagoen vocoderra aurkeztu zen artikulu honetan (Erro et al., 2011). HNM ereduak eskaintzen dituen abantailak aprobetxatuz (Stylianou, 1996), ahotsa parametrizatzen du bi korronte bidez: f0 eta espektroa. Vocoder honek oso emaitza onak lortzen ditu sintesian, kitzikapenari lotutako parametrorik erabili ez arren. Artikulu honetan vocoder hau parametro bat bakarrik gehituz, ahostun frekuentzia handienarena, hain zuzen ere (*Maximum Voiced Frequency*, MVF), hobetu daitekeela erakusten dugu. HNM lan-inguruan MVF espektroa bi zatitan banatzen duen frekuentzia da, beheko zatia harmonikoa edo ahostuna eta goikoa zaratatsua edo ahoskabea izanda. Nolabait, hau kitzikapen-eredu gisa erabiltzea bi bandako eredua erabiltzea bezalakoa da (Drugman et al., 2009)(Kim et al., 2006)(Silen et al., 2009), nahiz eta kasu hauetan MVFrekin erlazionatuta dauden parametrizazio eta berreraiketa prozedurek seinale osoa erabiltzen duten eta ez bakarrik kitzikapena. Vocoderren aurreko bertsioetan (Erro et al., 2010) MVF konstantea erabili zen. Hurrengo bertsioan MVF seinalearen energiaren arabera aldatzen bazen emaitza hobeak lortzen zirela konturatu ginen (Erro et al., 2011). Estrategia honek agertzen ziren metalezko zaratak leuntzen ditu. MVF energia jakinda kalkulatu zitekeenez, sistemak bi parametro-korronte behar zituen bakarrik eta beste aztergailu batzuekin bateragarria egin zitekeen. Hala ere, sistema hau ez zen inoiz konparatu hiru parametro-korronte (MVF barne) erabiltzen dituen sistema batekin. Artikulu honetan hurrengo urrats hau ematen da: ahots-sintesiko lan-inguru batean MVF estimatzeak eta modelatzeak merezi duen aztertzen da. Lortzen diren emaitzek HNM ereduan oinarritutako vocoderra kalitate handiko beste vocoder batzuekin konparatuta aukera ona dela erakusten dute.

Bigarren atalean MVF estimatzeko teknikak eta batez ere gure HNM ereduan oinarritutako vocoderrean integratu dena deskribatzen dira. Hirugarren atalean vocoder hau zehazten da. Bere ebaluazio pertzeptuala eta honen emaitza laugarren atalean komentatzen dira. Amaitzeko, ondorioak bosgarren atalean ateratzen dira.

2. MVF estimatzeko metodoak

HNM ereduak (Stylianou, 1996) ahots-seinalea bi osagaiz osatuta dagoela suposatzen du: alde batetik osagai harmonikoa, ahots-kordek dardara egiten dutenean sortzen dena eta denbora tarte txikietan periodikoa dela hausnar daitekeena eta beste aldetik osagai zaratatsua, seinalearen beste guztia bateratzen duena. Zati ahostunetan, MVF baloreak bitan zatitzen du espektroa: beheko banda harmonikoa eta goiko banda zaratatsua. Normalean 4kHz erabiltzen da MVF balore konstante bezala eta emaitzak onak izaten dira (Stylianou, 1996) (Drugman et al., 2009). Gure vocoderrak ere horrela erabiltzen zuen lehenengo bertsioan (Erro et al., 2010). Gero, 0. koefiziente cepstralaren arabera aldatzen zuen MVFa erabili zen (Erro et al., 2011). Koefiziente honek energiaren informazioa ematen du. Estrategia sinple honekin lehen bertsioan energia baxuko zatietan (esaldiaren amaieran adibidez) agertzen ziren zarata metaliko batzuk kentzea lortu zen. Hala ere, eta bi bandako eredua nahiko sinplea izan arren, MVF beste korronte bat balitz bezala tratatzeak seinalearen parametrizazio hobea eragingo zuen. Gainera, f0 detekzioan erroreak egiten direnean (normalean zati ahostunak hasten edo bukatzen direnean) burrunba entzuten da seinalean eta metodo honekin hau ere konponduko zen. Errore hau egiten denean normalean MVF txikiegiak estimatzen dira, beraz trama hori ea ahoskabea balitz bezala tratatuko da eta f0 errorea ez da hainbeste entzungo.

Orain arte MVF kalkulatzeko metodo ugari erabili da. Orokorrean ideia nagusia denbora laburreko espektroan gailurren harmonikotasun-gradua neurtzea da. Hau estimatzeko metodo batzuek espektroko gailurren eta sinusoide baten espektroaren arteko distortsioa kalkulatzen dute (Griffin eta Lim, 1988)(McAuley eta Quatieri, 1995). Beste batzuek gailurren anplitudea beraien bailarekin konparatuta hartzen dute kontuan (Erro et al., 2011)(Hermus et al., 2007). Aipatzekoa da autokorrelazio metodoa ere (Erro et al., 2010), f0 detekzio-zorroztasunarekiko eta formanteen posiziorekiko sentikorregia izan arren.

Gure proposamena sinusoide antzekotasun neurrian (*Sinusoidal Likeness Measure*, SLM) oinarritzen da. Hau erabiltzen da musika arloan, gailur espektralak sailkatzeko (Rodet, 1997). Oraingo tramaren N puntuko espektro konplexua, *S[k]*, kalkulatzen da, 3 periodoko Hanning leihoa erabiliz. Beraz, f0 jakin behar da. N 4L baino handiagoa den lehenengo biren potentzia da, 4 zeroen faktore bateragarria da eta L tramaren luzera da. Anplitude espektroaren gailurren frekuentziak $\{f_i\}$ kalkulatzen dira maximoen inguruan hurbilketa parabolikoa aplikatuz, eta haien SLMa kalkulatzen da korrelazio gurutzatu lokala erabiliz (Rodet, 1997):

$$L_i = \frac{\left|\sum S[k]\cdot W_i^*[k]\right|}{\sqrt{\sum |S[k]|^2 \cdot \sum |W_i[k]|^2}} \quad \forall k, \left|k\frac{f_s}{N} - f_i\right| < \frac{f_0}{2} \qquad (1)$$

W_i azterketa leihoaren Fourier transformatua bider f_i kosinua da, * konjugatu konplexua da, f_0 pitch lokala da eta f_s laginketa-maiztasuna da. W_i ondo hurbildu daiteke espresio analitikoen bidez. SLM 0tik 1tera doa. Sinusoide hutsen SLMek 1 inguruko balioa daukate eta balio txikiagoek zarata dagoela edo analisi-trama denboran aldatzen den sinusoideak daudela adierazten dute. Analisi-bandako gailur guztietarako SLM kalkulatu eta gero, MVF gailur bakoitzean jartzen denean egiten den errorea kalkulatzen da:

$$\varepsilon_i^2 = \frac{1}{I}\left(\sum_{j<i}\left(1-L_j\right)^2 + \sum_{j\geq i}\left(\max\left\{L_j,\lambda\right\}-\lambda\right)^2\right) \quad (2)$$

I espektroko gailur-kopurua da eta λ ahostun atalasea dela suposa daiteke. Gure lanean 0.85 balioa zeukan. Horrelako errore-neurketa benetako SLM inguratzaile eta MVFa oraingo gailurrean ezartzen duen bi bandako inguratzaile idealaren arteko distantzia balitz bezala ulertu daiteke. Errore-funtzio honetako minimo lokalak (normalean bat baino gehiago dago) MVF hautagai gisa hartzen dira. Azken erabakia MVF

$$C\left(\left\{f_{t,i(t)}\right\}_{t=1}^{T}\right) = \sum_{t=1}^{T}\varepsilon_{t,i(t)}^2 + \gamma\sum_{t=2}^{T}\left(f_{t,i(t)} - f_{t-1,i(t-1)}\right)^2 \quad (3)$$

ibilbidean (denboran) $\{f_{t,i(t)}\}$ Viterbi bilaketa eginez hartzen da:

t denbora unea da, *i(t)* *t* momentuan kontuan hartzen ari den gailurraren indizea da eta γ pisu faktorea da. Gure frogetan emaitza onak lortu ziren γ=5.10^{-4}.r/f_s^2, r trama abiadura izanda). Trama ahoskabeetan, hautagai bakar bat hartzen da kontuan 0Hz balioarekin.

Deskribatutako metodoa bi parametro bidez ajusta daiteke: λ eta γ. MVFa ez da ahotsaren ezaugarri fisikoa, ahotsaren eredu sinplifikatua erabiltzen denean sortzen den balioa baizik. Horregatik ez dago MVFarekin etiketatutako datu-baserik. Beraz, parametroen optimizazioa eskuz egin da entzumen froga informalen bitartez.

3. Vocoderraren deskripzioa

Vocoderraren bertsio hobetuak hiru korrontetan parametrizatzen du ahotsa: f_0-n, MVFn eta espektroan. Hurrengo paragrafoetan nola erauzten diren parametro hauek eta parametro hauetatik abiatuta nola berreraikitzen den seinalea azaltzen da. Deskripzio zehatzagoa irakur daiteke Erro et al., (2011) artikuluan.

Bai f_0 eta MVF eskalarrak dira: f_0 pitch kalkulatzeko baliogarria den edozein algoritmoren bidez kalkula daiteke (Luengo et al., 2007) eta MVFa 2. atalean aurkeztu den metodoaren bidez kalkulatzen da. Espektroa irudikatzeko *p+1* koefiziente cepstralen bidez irudikatzen da (energiakoa barne). Trama ahostunak eta ahoskabeak era ezberdinean tratatzen dira koefiziente cepstralak kalkulatzeko. Trama ahostuna bada, karratu txikienen optimizazioan oinarritzen den analisi harmonikoa egiten da (Stylianou, 1996) banda osoan f_0-ren multiploa diren frekuentzietan harmonikoen anplitudea lortzeko. Anplitude hauek benetako inguratzaile espektralaren lagin gisa hartzen dira, frekuentzia handienetan ere, harmoniko/zarata ratio txikia bada ere. Trama ahoskabeak FFT bidez aztertzen dira. Espektroaren irudikapena homogeneizatzeko, trama ahostunetan anplitude harmonikoek definitzen duten inguratzailea anplitudean normalizatzen da eta gero FFT erresoluzioarekin lagintzen da interpolazioaren bidez (Erro et al., 2011). Analisiaren azken pausuan anplitude espektrotik abiatuta koefiziente cepstralak erauzten dira: lehengo ohizko cepstruma kalkulatzen da eta ondoren (Tokuda et al., 1994) artikuluan deskribatzen den errekurtsioa erabiltzen da Mel eskala aplikatzeko.

Seinalea berreraikitzeko eta trama bakoitzeko laginak parametroetatik sortu eta gero, teilakapen eta gehitzea teknikak (overlap-add, OLA) erabiltzen dira. Trama bakoitza HNM sintesi-prozedurak erabiliz sortzen da. Lehenengo zati zaratatsua sortzen da (zati hau bai trama ahostunetan bai ahoskabeetan dago): zarataren modulua cepstral inguratzailea laginduz lortzen da eta faseari balore ausazkoak ematen zaizkio. Trama ahoskabeak zarata espektroari alderantzizko FFT aplikatuz lortzen dira. Trama ahostunetan zarata espektroa MVF mozketa frekuentzia daukan goi-pasako iragazkiaz biderkatzen da eta gero alderantzizko FFT aplikatzen da. Ondoren zati harmonikoa sortzen da denboran. Anplitude harmonikoak inguratzaile cepstrala f_0 multiploetan laginduz lortzen dira; faseak kalkulatzeko fase minimoaren hurbilketa erabiltzen da. Tramen artean fasearen koherentzia mantentzeko fasean frekuentziarekiko termino lineala sartzen da.

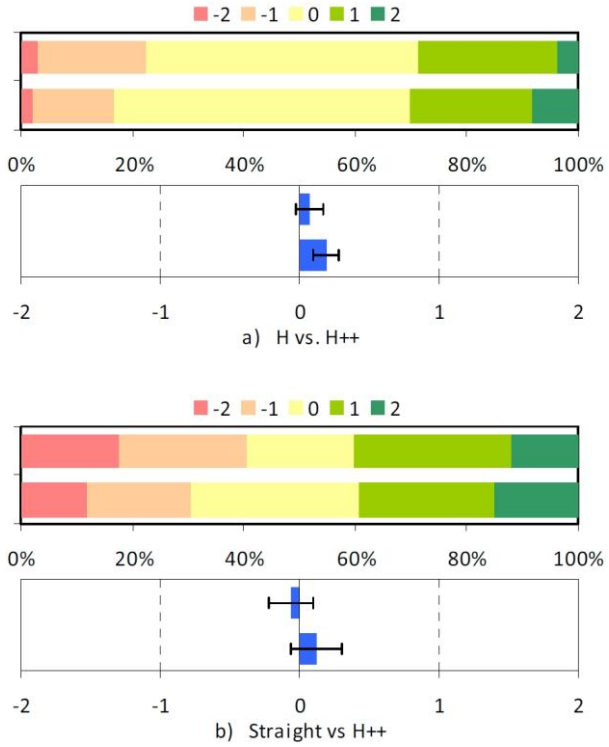
4. Ebaluazioa

MVF estimatzeko metodoa ebaluatzeko datu-base aproposa ez dago, beraz froga subjektiboak erabili dira 3 korronteko vocoder osoa ebaluatzeko. Honen bersintesi gaitasuna informalki ebaluatu zen eta orduan vocoderra sintesi lan-inguruan frogatu zen, HTS erabiliz. HTS paketea HTS working group 2002an garatutako software libreko programa da (http://hts.sp.nitech.ac.jp/). HTS 2.1.1 bertsioak hizlariaren menpeko eta hizlariari egokitzeko sistemak entrenatzeko demo programak ematen ditu. Vocoder-metodo biak dauzka: oinarrizkoa, mel-cepstral analisian eta kitzikapen eredu sinplean oinarrituta eta Straight-en oinarritutakoa (Zen et al., 2007), oso emaitza onak ematen dituena.

HMMetan oinarritutako sintesi-sistema eraiki zen HTS eta AhoTTS (Erro et al., 2010) sistemaren modulu linguistikoa konbinatuz. Ahots sintetikoa hiru vocoder ezberdin erabiliz egin da: artikulu honetan deskribatu den HNM ereduan oinarritutako vocoder hobetua (H++ sinboloarekin irudikatzen dena), vocoder honen aurreko bertsioa (H sinboloarekin irudikatzen dena) eta Straight. Hiruretan koefiziente cepstral kopuru bera erabili da (39+energia), baina kitzikapenaren parametro kopurua ezberdina da kasu guztietan: H-n bat ere ez, H++-n parametro bat bakarrik (MVF) eta Straight-en 5 banda ez-periodikotasun.

Bi datu-base erabili dira ebaluaketan erabili diren ahotsak sortzeko: lehenengoa emakume batek euskara batuan ahoskatutako 2K esaldi labur (bi hizketa ordu baino gehiago) eta bigarrena gizon batek erdaraz ahoskatutako 1.2K esaldi dauka (bi hizketa ordu). Biak estilo neutroan grabatu ziren.

Bi konparaketa egin dira CMOS (*Comparative Mean Opinion Score*) test batekin. Lehenengoan H eta H++ konparatu dira eta bigarrenean Straight eta H++. Ausazko 12 esaldi paretan, ausaz aukeratutako entzuleek haien zaletasuna azaldu behar zuten 1-5 eskalan: "askoz nahiago dut A seinalea B baino", "pixka bat nahiago dut A seinalea B baino", "biak berdin zaizkit", "pixka bat nahiago dut B seinalea A baino", "askoz nahiago dut B seinalea A baino". Ahots naturalaren grabazio batzuk ere sartu ziren erreferentzia gisa erabiltzeko. Eskalako puntu bakoitzari balio numerikoa eman zitzaion (-2tik 2ra) eta CMOS balio finala entzuleek eman zituzten CMOS balio guztien bataz besteko balorea kalkulatuz lortu zen. Bi frogetan 2 balioa eman zitzaion H++ zaletasunari. Lehenengo frogan 45 entzulek eta bigarrenean 30 entzulek hartu zuten parte.



1. Irudia: *Scoreen distribuzioa eta CMOS bi metodo pare eta ahots ezberdinetarako. (a) H vs. H++; (b) Straight vs. H++. Goian erdarazko gizonezko ahotsa eta behean euskarazko emakumezko ahotsa*

1a irudian ikus daitekeenez, H++ vocoderra H vocoderra baino pixka bat nahiago dute entzuleek (30% H++ vs. 20% H). Hala ere, ezberdintasunak ez dira erraz nabaritzen. Bi ondorio atera daitezke: MVF~~a~~ modelatzeak sintesian ahots- kalitate hobea lortzen du, beraz 2. atalean aurkeztu den MVF~~a~~ kalkulatzeko metodoa HNM vocoderrean sartzeak merezi du; bestalde, H vocoderrean erabiltzen den hurbilketa (MVF energiaren arabera kalkulatzea) aplikazio praktikoetan nahikoa da. Ezberdintasunak zailagoak dira entzuteko gizonezko ahotsean.

1b irudian aurkezten diren zaletasun-distribuzioetan bi metodoen arteko ezberdintasunak ondo nabaritu direla ikus daiteke, bataz besteko zaletasuna argi ez izan arren. Nahiz eta entzule kopuru handia izan, emaitzak ez dira esangarriak ondorioak atera ahal izateko. % 95 konfiantza tarteek 0a (zaletasun falta) barne daukate bi ahotsetarako. Hala ere, proposatzen den metodoak Straight-ek baino score hobeak lortzen ditu emakumezko ahotserako.

Beraz, artikulu honetan aurkeztu den vocoderra interesgarria da Straight-arekiko aukera gisa erabiltzeko, ahots batzuetarako behintzat. Vocoderra liberatu nahi da hainbat arinen.

5. Ondorioak

Artikulu honetan ahostun frekuentzia maximoa (espektroa banda harmoniko batean eta zarata-banda batean banatzen duena) estimatzearen eta modelatzearen eragina aztertu da. Eraiki den sistemak aurreko bertsioak baino emaitza pixka bat hobeak lortzen ditu. Azterketa egiteko erabili den ahots batean hobekuntzak nabariagoak izan dira, eta kasu honetarako Straight vocoderrak baino emaitza hobeak lortzen ditu. Beraz, erabili daiteke kalitate handiko vocoderrak eraikitzeko.

Hurrengo lanetan, MVF estimatzeko, atalase-balioen pitch-balioarekiko menpekotasuna kontuan hartu beharko da. Vocoderrari dagokionez, analisia hobetzen ari da efizienteagoa egiteko; horrela denbora errealeko aplikazioetan erabilgarria izango da. Vocoderra denbora asko barik publikoa izango da.

6. Eskerronak

Lan hau UPV/EHUko laguntzarekin (Ayuda de especialización de doctores), Espainiako Zientzia eta Berrikuntza Ministerioko laguntzarekin (Buceador proiektua, TEC2009-14094-C04-02) eta Eusko Jaurlaritzako laguntzarekin (Berbatek, IE09-262) egin da.

7. Aipamenak

T. Drugman, G. Wilfart, T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis", Proc. Interspeech, pp.1779-1782, 2009.

D. Erro, I. Sainz, E. Navas, I. Hernaez, "HNM-based MFCC+F0 extractor applied to statistical speech synthesis", Proc. ICASSP, pp. 4728-4731, 2011.

D. Erro, I. Sainz, I. Saratxaga, E. Navas, I. Hernaez, "MFCC+F0 extraction and waveform reconstruction using HNM: preliminary results in an HMM-based synthesizer", Proc. FALA 2010 (VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop), pp. 29-32, 2010.

D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, I. Hernáez. HMM-based Speech

Synthesis in Basque Language using HTS. Proc. FALA 2010 (VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop), pp. 67-70, Vigo, Spain, 2010.
X. Gonzalvo, J.C. Socoro, I. Iriondo, C. Monzo, E. Martinez, "Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish", Proc. 6th ISCA Speech Synthesis Workshop, pp. 362–367, 2007.
D.W. Griffin, J.S. Lim, "Multiband Excitation Vocoder", IEEE Trans. Acoust., Speech & Sig. Proc., vol. 36(8), pp. 1223-1235, 1988.
K. Hermus, H. van Hamme, S. Irhimeh, "Estimation of the voicing cut-off frequency contour based on a cumulative harmonicity score", IEEE. Sig. Proc. Letters, vol. 14(11), pp. 820-823, 2007.
A. Hunt, A. Black, "Unit selection in a concatenative speech synthesis system using a large speech data-base", Proc. ICASSP, pp. 373–376, 1996.
H. Kawahara, "Straight, exploration of the other aspect of Vocoder: perceptually isomorphic decomposition of speech sounds", Acoustic Science and Technology, vol.27, no.6, pp.349-353, 2006.
S.J. Kim, J.J. Kim, M. Hahn, "HMM-based Korean speech synthesis system for hand-held devices", IEEE Trans. Consumer Electronics, vol. 52(4), pp. 1384-1390, 2006.
I. Luengo, I. Saratxaga, E. Navas, I. Hernaez, J. Sanchez, I. Sainz, "Evaluation of pitch detection algorithms under real conditions", Proc. ICASSP, pp. 1057-1060, 2007.
R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling", Proc. 6th ISCA Speech Synthesis Workshop, pp.131-136, 2007.
R. McAulay and T. Quatieri, "Sinusoidal Coding", chapter in "Speech Coding and Synthesis", Elsevier, pp.121-173, 1995.
F. Mendez, L. Docio, M. Arza, F. Campillo, "The Albayzin 2010 text–to–speech evaluation", Proc. FALA, pp. 317-340, 2010.
T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, P. Alku, "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering", IEEE Trans. Audio, Speech, & Language Processing, vol. 19(1), pp. 153-165, 2011.
X. Rodet, "Musical Sound Signals Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models", Applied Sig. Proc., vol. 4, pp. 131-141, 1997.
H. Silen, E. Helander, J. Nurminen, M. Gabbouj, "Parameterization of vocal fry in HMM-based speech synthesis", Proc. Interspeech, pp. 1775-1778, 2009.
Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification", PhD thesis, École Nationale Supèrieure des Télécommunications, Paris, 1996.
K. Tokuda, T. Kobayashi, T. Masuko, S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation", Proc. Int. Conf. Spoken Language Processing, vol.3, pp.1043-1046, 1994.
T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Mixed excitation for HMM-based speech synthesis", Proc. Eurospeech, pp.2263–2266, 2001.
H. Zen, K. Tokuda, A.W. Black, "Statistical parametric speech synthesis", Speech Communication, vol.51, no.11, pp.1039-1064, 2009.
H. Zen, T. Toda, M. Nakamura, K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005", IEICE Trans. Inf. Syst. E90-D (1), pp.325–333, 2007.