# MFCC+F0 erauzketa eta ahots-uhina berreraikitzea HNM erabiliz: HMM sintesi-sistema batean lortutako aurretiko emaitzak

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

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

**Abstract**

The most widespread techniques for speech synthesis and voice conversion are currently based on probabilistic frameworks. Particularly, Hidden Markov Models (HMMs) play a relevant role in speech synthesis, whereas Gaussian Mixture Models (GMMs) are almost standard in voice conversion. Consequently, in both cases the performance of the systems is limited by three main factors: 1) the suitability of the statistical models; 2) the over-smoothing phenomenon; 3) the accuracy of the underlying speech parameterization and reconstruction method. This paper focuses on the third issue, still open at present: translating speech frames into parameter vectors with good properties for the mentioned statistical frameworks, and reconstructing waveforms properly. The proposal presented in this paper uses the Harmonics plus Noise Model (HNM) to extract MFCC+f0 and reconstruct speech frames from them. The results of a perceptual evaluation show that the tool is valid for state-of-the-art HMM-based speech synthesis systems.

**Laburpena**

Gaur egun hizketa sintetizatzeko eta bihurtzeko teknikarik hedatuenak probabilitateetan oinarritzen dira. Bereiziki Markoven Ezkutuko Ereduak (Hidden Markov Models, HMM) hizketa-sintesian aipagarriak dira eta Nahasketa Gaussiarren Ereduak (Gaussian Mixture Models , GMM) ia estandarrak dira ahots-bihurketan. Ondorioz, kasu bietan sistemaren errendimendua hiru faktore hauek mugatuta dago: 1) eredu estatistikoen komenigarritasuna; 2) gehiegizko leunketa; 3) ahotsa parametrizatzeko eta berreraikitzeko erabiltzen den metodoaren zehaztasuna. Artikulu honetan deskribatzen den lana gaur egun ebatzi gabe dagoen hirugarren arazoaz arduratzen da, hau da, ahotsa ezaugarri estatistiko onargarrietako parametro nola bihurtu eta ahots-uhina nola berreraiki. Proposatzen den metodoak Harmonikoak gehi Zarata Eredua (Harmonics plus Noise Model, HNM) erabiltzen du MFCC+f0 erauzteko eta hauetatik abiatuta, ahots-tramak berreraikitzeko. Pertzepzio-ebaluaketaren emaitzek erakusten dute metodoa egokia dela HMMetan oinarritutako gaur egungo sintesi-sistemetan erabiltzeko.

**Keywords:** speech parameterization, statistical parametric speech synthesis, voice conversion, harmonics plus noise model

**Hitz gakoak:** Ahots-parametrizazioa, ahots-sintesi parametriko eta estatistikoa, ahots-bihurketa, harmonikoak eta zarata eredua

## 1. Sarrera

Ahots-parametrizazioa eta berreraikuntza gaurkotasun handiko gaiak dira, batez ere HMMetan oinarritutako hizketa-sintesi-sistemek (Zen et al., 2009) (Yamagishi et al., 2009) eta GMMetan oinarritutako ahots-bihurketa-sistemek garapen handia izan dutelako (Stylianou et al., 1998)(Kain, 2001)(Toda et al., 2007)(Erro et al., 2010). Lan esparru estatistiko hauek ahotsa ezaugarri onak dauzkan bektore-multzo erabilerraz bihurtu behar dute. Horregatik ahots-sintesi eta bihurketan espektroa modelatzeko MFCC koefizienteak erabiltzen dira (Zen et al., 2009)( Toda et al., 2007). Parametro hauek dauzkaten abantailen artean hurrengoa aipatu behar da: kobariantza-matrize diagonalak erabiltzen uzten dutela, bektore osatzaileak korrelazio gabekoak direlako. Ahots-bihurketan bestelako parametroak ere erabiltzen dira, Line Spectral Frequencies (LSF), hain zuzen ere (Kain, 2001)( Erro et al., 2010). Hala ere, ahots-seinaletik parametro bektoreak erauzteko modu bakarra ez dago eta are gutxiago ahotsa berreraikitzeko prozedura bakarra. Ahotsa kodetzeko eta deskodetzeko metodoek ikertzeko aukerak eskaintzen dituzte, bai ahotsetik parametro erauzketak eta bai parametroetatik ahotsa berreraikuntzak sistemen funtzionamenduan eragin handia daukate. Arazo hau larriagoa da ahots-sintesian ahots-bihurketan baino, ahots-bihurketan jatorrizko hizlariaren ahots naturala erabilgarria delako eta abiapuntu bezala erabil daitekeelako. Horregatik, artikulu hau eta hemen azaldutako lana sintesirako daude pentsatuta.

HMMetan oinarritutako sintesi-sistemen kasu berezian, parametrizazio modu anitz erabili dira azken hamabost urteetan. HTS HMM bidezko sintesi-sistema publikoa da (Zen et al., 2007), HTK paketean (Young et al., 2006) oinarrituta dago eta Nitech-en sortu zen. Bere oinarrizko inplementazioan espektroa MFCC koefizienteen bidez irudikatzen da eta Mel-hedatu analisi cepstrala erabiltzen da koefiziente hauek lortzeko (Tokuda et al., 1994). Pultsu/zarata oso kitzikapen sinplea erabiltzen da $f_0$-n oinarrituta (Yoshimura et al., 1999). Hurrengo bertsioak kitzikapen nahastu sofistikatuagoa erabiltzen du (Yoshimura et al., 2001)(Gonzalvo et al., 2007). Maiak eta lankideek (2007) are sofistikatuagoa den kitzikapen entrenagarria

erabiltzen zuten, zarata eta pultsuetarako egoeraren menpeko iragazkietan oinarritzen zena. Lan berri batean, Drugmanek eta bere lankideek (2009) bi bandako kitzikapen nahasia erabiltzen zuten. Kitzikapen honetan goiko banda zarataz osatuta dago eta beheko banda osagai nagusizko analisiaren bidez aukeratzen diren uhin deterministek osatzen dute. Hemptinnek (2006) eta Banosek eta lankideek (2008) seinale bera (eta ez bakarrik kitzikapena) modelatzen zuten harmonikoak eta zarata bidez. Bi lan hauetan entrenatzeko erabiltzen diren parametroak aurresate linealean oinarritzen ziren. Beste lan batzuek iturri glotala eta ahots-traktua erabiltzen dute, kitzikapena eta espektroa erabili beharrean (Cabral et al., 2008)(Raitio et al., 2010)(Lanchantin et al., 2010). Parametroen erauzketa eta eredu estatistikoen eraikuntza integratzeko saiakerak ere egin dira (Toda eta Tokuda, 2008). Ebazpenik erabiliena Straight-ek erabiltzen duena da. Straight kalitate handiko vocoderra da eta seinalea bi osagaiaz konposatuta dagoela suposatzen du: alde batetik inguratzaile espektrala eta beste aldetik $f_0$ eta inguratzaile ez-periodikoa delakoa (Kawahara, 2006). Straight sistemak sortzen duen irteera beste parametro egokiago bihurtzen da, normalean MFCC eta banda ez-periodikotasunak (Zen et al., 2007). Hala ere Straight software itxia da.

Artikulu honek, seinale-tramak abiatuta, MFCC+f0 (eta alderantziz) erauzten duen tresna aurkezten du. Tresnak HNM eredua erabiltzen du seinaleak irudikatzeko (Stylianou, 1996) eta HTS integratzeko pentsatu da. Erabilitako metodoak ezaugarri interesgarri hauek dauzka:

- Orden handiko MFCCak erauzteko ahalmena dauka.
- Kitzikapenarekin erlazionatzen den parametro bakarra $f_0$ da.
- Bersintesian kalitate handia lortzen du.
- Ahots-aldaketa eta manipulazio batzuk egin daitezke.
- Seinalea berreraikitzeko prozedura oso efizientea da.

Tresna hau sintesian ebaluatzeko egindako froga pertzeptualek Straighten mailan dagoela erakusten dute. Tresna askatuko dugu hurrengo hilabeteetan. Hurrengo atalean metodoa deskribatuko da, 3. atalean egindako ebaluaketaren emaitzak aurkeztuko dira eta 4. atalean ondorioak aterako dira.

## 2. Metodoaren deskribapena

Metodoa Larochek eta bere lankideek (1993) proposatutako ideian oinarrituta dago. Seinalea zati harmonikoan eta estokastikoan banatzen da. Zati harmonikoak ahots-kordek dardara egiten dutenean sortzen den seinale lokalki periodikoa antzematen du. Harmonikoki erlazionatutako sinusoideek egiten dute zati honen eredua. Zati estoskastikoak zati harmonikoak antzeman ezin duen fenomeno guztiak irudikatzen ditu. Normalean forma emateko iragazki batetik pasa den zarata gaussiar zuria erabiltzen da zati honen eredua eraikitzeko.

$$s(t) = \sum_i A_i(t) \cdot \cos\left(2\pi i f_0(t) t + \varphi_i(t)\right) + e(t) \quad (1)$$

Ahots-seinalearen eredu honek eta dagozkion algoritmo eta metodoek (Stylianou, 1996) ahots-seinalea aztertzeko, aldatzeko eta bererraikitzeko egokia den kalitate handiko parametrizazioa hornitzen dute. Metodo honen beste inplementazioa trama maiztasun egonkorra erabiltzeko egokiagoa da (Erro et al., 2007). Hala ere, halako parametrizazioa oso zaila da eremu estatistikoan erabiltzeko (Banos et al., 2008), beste arrazoien artean honako hauengatik:

- Analisi banda barruan dauden harmoniko kopurua aldatzen da $f_0$-ren arabera.
- Ateratzen diren parametro kopurua handia da ($f_0$=1000Hz, 0 eta 5kHz artean 50 harmoniko daudela esan nahi du, bakoitzak bere anplitude eta fasearekin).
- Anplitudeak eta faseak asko aldatzen dira $f_0$-ren arabera.

Horregatik, eredua ez da egokia lan ingurune estatistikoan ahotsa zuzenean parametrizatzeko, beste parametro mota batzuk erauzteko laguntza bezala erabili daitekeen arren (Hemptinne, 2006)(Banos et al., 2008). Hurrengo azpiataletan analisia egiteko eta seinalea berreraikitzeko metodoak azaltzen dira.

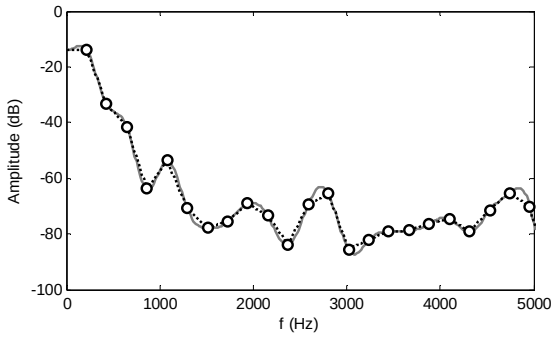
### 2.1. Parametroak erauzi

Analisia egiterakoan sistemak, sarrera seinalea, analisiaren trama maiztasuna eta parametrizazioaren ordena jakinda, $f_0$ balorea eta MFCC bektorea kalkulatzen du trama bakoitzerako.

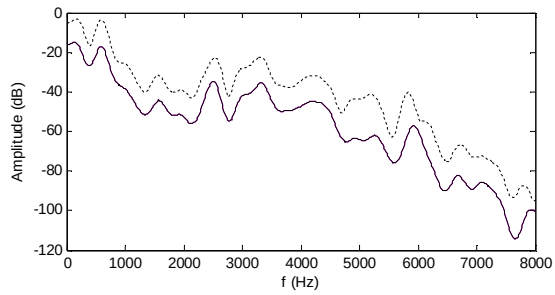
Analisian lehen urratsa pitch balorea kalkulatzea da. Kasu honetan autokorrelazioan oinarritzen den metodoaren aldaera bat erabiltzen da (Boersma, 1993) $f_0$ lokala kalkulatzeko eta oraingo trama ahostuna edo ahoskabea den jakiteko. Jatorrizko metodoan egindako aldaketak estimazioa zehatzagoa izateko egin dira.

MFFC parametroak kalkulatzeko trama ahostunak eta ahoskabeak era ezberdinean erabiltzen dira. Sarrerako trama ahostuna bada analisi harmoniko arrunta (karratu txikien optimizazioan oinarrituta (Stylianou, 1996)) egiten da analisi-banda osoan. Horrela harmonikoen log-anplitudeak $f_0$ multiploetan lortzen dira. Anplitude hauek inguratzaile espektralaren laginak bezala interpretatu daitezke. Frekuentzia handietan ere (Nyquist-en frekuentziatik hurbil) analisi harmonikoak teorikoki inguratzailearen lagin baliagarriak kalkulatzen ditu. Fourier transformatu azkarra (Fast Fourier Transform, FFT) erabiltzen da trama ahoskabeak aztertzeko. Nahi bada espektroa banda batzuetan leundu daiteke. Bi irteera motak

homogeneizatzeko anplitude harmonikoek irudikatzen duten inguratzailea FFTak daukan erresoluzioarekin lagintzen da berriro, interpolazioa erabiliz. Egindako ikerketa lan batzuek (Banga et al., 2001)(Erro et al., 2007) interpolazio lineala pitch aldaketa eta beste aplikazio baztuk egiteko zehatza dela erakutsi arren, lan honetan sinc bidezko interpolazioa erabili da analisia konsistenteagoa egiteko (Ikusi 1. Irudia xehetasunak lortzeko). $f_0$ baino txikiagoak diren frekuentzietan informazio espektrala fidagarria ez denez, interpolazioa egin baino lehen harmoniko artifiziala sartzen da 0 Hz. Harmoniko honek $f_0$-koak daukan anplitude bera dauka. Banga eta bere lankideek (2001) erabili zuten antzeko estrategia eta emaitza pertzeptual onak lortu zituzten. Lortzen den inguratzaile espektrala eta Straight-en bidez kalkulatzen dena oso antzekoak izan behar dira (Ikusi 2. Irudia) eta, beraz, abantaila posible berberak ditu, batez ere orden handiko MFCC parametroak estimatzen uztea.




1.Irudia: *Anplitude sinc bidezko interpolazioa (lerro jarraitua) vs. interpolazio lineala (lerro ez jarraitua).*




2. Irudia: *Proposatutako metodoak lortzen duen espektroa (lerro jarraitua) eta Straight-en espektroa (lerro ez jarraitua) trama ahostun baterako*

Anplitude-espektroa normalizatzen da $f_0^{-½}$ biderka faktorea erabiliz (trama ahoskabeetan $f_0$-k $f_s/L$ balioa hartzen du, $f_s$ laginketa-maiztasuna eta L analisi-leihoaren luzera izanda). Normalizazio hau beharrezkoa da anplitudeak daukan $f_0$-ren menpekotasuna ezabatzeko. Honela neurtutakoa ez diren f0 baloreak erabili daitezke seinalea bersintetizatzeko. Bi seinalek energia eta inguratzaile espektral bera badaukate, haien pitch-ari proportzionalak diren anplitudeak izango dituzte. Azalpena erraza da: banda-zabalera finkatzen bada, $f_0$ balore handietarako energia harmoniko gutxiagok eman behar dituzte, beraz, haien anplitudeak handiagoak izan behar dute.

Analisiaren azken urratsean koefiziente cepstralak anplitude espektrotik kalkulatzen dira. Lehendabizi ohiko cepstrum-a kalkulatzen da log-anplitude espektroaren Fourier alderantzizko transformatua bezala. Orduan dimentsioa murrizten da eta parametrizazio cepstralaren warping faktorea transformatzen da, Tokuda eta lankideek (1994) erabiltzen duten Mel eskala egokitzeko. Espektrotik abiatuta MFCC-ak kalkulatzeko beste metodo batzuk ere frogatu dira (Cappé et al., 1995), baina froga informalek deskribatutako moduak emaitza hobeak ematen zituela erakutsi zuten.

## 2.2. Ahots-uhina berreraiki

Lehenengo urratsa seinalearen zati zaratatsua sortzean datza, zati hau bai soinu ahostunek bai ahoskabeek baitaukate. MFCC koefizienteak erabiltzen dira FFT espektroa berreraikitzeko eta alderantzizko FFTak ematen digu zarata. MFCC inguratzailea lagintzen da erresoluzio egoki batekin (100 Hz) eta interpolazio lineala erabiltzen da FFTak behar duen erresoluzioa lortzeko. (fs/L)½ faktorea erabiltzen da inguratzailearen normalizazioa orekatzeko (kasu honetan L FFT luzera da) eta FFT modulua lortzen da. Fasea ausazkoki sortzen da $[-\pi, \pi)$ tartean distribuzio uniformea suposatuta.

Sortu behar den trama ahoskabea bada, trama sintetikoa kalkulatu den zaratak osatzen du. Bestela, zarata goi-pasako iragazkiaz iragazten da frekuentzian (alderantzizko FFTa aplikatu baino lehen). Iragazki honetan ahostun frekuentzia maximoa erabiltzen da (5 kHz balio egokia dela erakutsi da (Erro et al., 2007)). Iragazkiaren forma konstantea da, hau da, zati zaratatsua ez da esplizitoki modelatzen, baina hala erabaki da honelako HNM ereduak lortu dituen emaitza onak ikusita (Stylianou, 2009). Gero zati harmonikoa sortzen da hurrengo prozedura jarraituz. MFCC inguratzailea lagintzen da eta $f_0^{½}$ faktorea erabiltzen da inguratzailearen normalizazioa orekatzeko, anplitudeak kalkulatzeko. Faseak fase minimoa erabiliz kalkulatzen dira (McAulay eta Quatieri, 1995). Are gehiago frekuentzian lineala den fase terminoa (Banos et al., 2008) gehitzen zaio trama bakoitzari alboko tramekin fase erlazio koherentea mantentzeko. 3.5kHz baino handiagoak diren harmonikoetan fase dispertsio artifiziala sartzen da, ahots sintetikoan agertzen den burrunba murrizteko. Seinale sintetiko naturalagoak lortzeko, beste fase-manipulazio teknikak batzuk ere frogatu dira (Ahmadi eta Spanias, 2001)(Sun et al., 1997), baina ez dituzte deskribatutako metodoak baino emaitza hobeak lortu.

Seinale sintetikoa berreraikitzeko leiho triangeluarreko overlap-add (OLA) erabiltzen da. Seinalearen adierazpena hau da:

$$s(kT+t)=\frac{T-t}{T}\cdot s^{(k)}(t)+\frac{t}{T}\cdot s^{(k+1)}(t-T)\ ,\quad 0\le t<T$$
$$s^{(k)}(t)=\sum_{i=1}^{I^{(k)}} A_i^{(k)}\cos\left(2\pi i f_0^{(k)}t+\varphi_i^{(k)}\right)+e^{(k)}(t) \qquad (2)$$

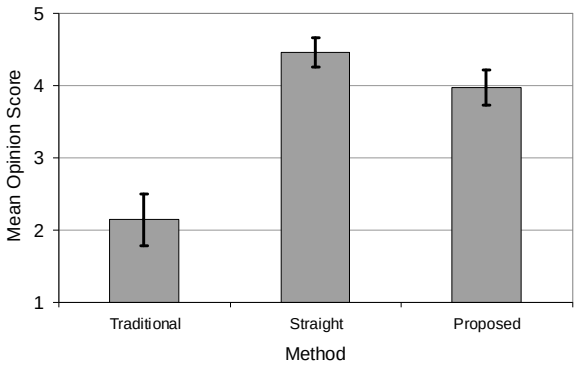
$\{A_i^{(k)}\}$, $\{\varphi_i^{(k)}\}$, eta $e^{(k)}(t)$ k. tramaren anplitudeak, faseak eta zarata dira eta $T$ tramen arteko distantzia da.

## 3. Atariko ebaluaketa

HTS (HMM-based speech synthesis system) deitzen den kode irekiko software paketea 2002tik eskuragai dago. Entrenamenduan ahots-seinale batzuen irudikapen parametrikoa eta eduki prosodiko eta fonetikoa deskribatzen duten etiketetatik abiatuta, HTS sistemak fonemen ezaugarri akustikoak eta iraupena modelatzen ditu, inguruaren menpeko HMMak erabiliz (CD-HMM). Sintesia egiterakoan sortu behar den seinaleari dagozkion etiketetatik abiatuta, HTS sistemak esaldi-HMMak sortzen ditu dagokion CD-HMMak konkatenatuz. Orduan, uhina sortzen du esaldi HMM ereduarekiko probabilitate handiagoa daukan bektore sekuentzia erabiliz.

Oraingo HTS distribuzioak hizlariaren menpeko eta hizlariari egokitutako sistemak entrenatu ditzake. HTS demoan bi metodo dago eskuragai parametrizazioa eta berreraikuntza egiteko: ohikoa (MFCC parametroak erabiltzen duena) eta Straight-en oinarritzen dena. 2. atalean proposatu den metodoa ebaluatzeko HTSn oinarritzen den sintesi-sistema eraiki zen eta seinale sintetikoen naturaltasuna neurtzeko MOS (Mean Opinion Score) frogak erabili ziren. Ebaluazioan erabili zen datu-baseak neska batek euskaraz eta estilo neutroan irakurri zituen 2.000 esaldi labur dauka (2 ordu gutxi gorabehera). Zazpi entzule boluntariok parte hartu zuten ebaluazioan. Ebaluatzaileek hiru bertsio ezberdinekin (ohikoa, Straight eta Proposatutakoa) sortu diren bost esaldi sintetiko entzun eta 1-5 arteko eskala batekin ebaluatu behar zituzten, 1 puntuaren esanahia “oso naturaltasun gutxikoa” eta 5 puntuaren esanahia “oso naturaltasun handia” izanez. Ohiko metodoan f0+25 MFCC koefiziente erabili ziren, Straight-en f0+40 MFCC+ 5 banda ez-periodikotasun eta, proposatzen den metodoan f0+40 MFCC koefiziente.

MOS emaitzak 3. irudian ikus daitezke %95 konfiantza tartearekin. Emaitzek erakusten dute proposatutako metodoa ohikoa baino askoz hobea dela. Straight-ek oraindik emaitza hobeak lortzen ditu, baina tartea txikiagoa da kasu honetan. Tarte txiki honen arrazoia osagai ez-periodikoa modelatzen datzala uste dugu. Hurrengo lanetan banda osoko zarata erabiltzen duten HNM aldaera ezberdinak (Erro et al., 2007) aztertuko dira.




3. Irudia: *MOS ebaluaketaren emaitzak, %95 konfiantza tarteak erabiliz*

## 4. Ondorioak

Artikulu honetan seinaletik MFCC eta $f_0$ erauzteko eta uhina berreraikitzeko metodoa aurkeztu da. Proposatutako metodoa HNM ereduan oinarrituta dago eta, HMM sintesi-sistemarik onenekin konparatuta, emaitza onak lortzen ditu. Artikulu honetan aukeztu diren atariko emaitzak ez daude Straight sistemak lortzen dituenetatik oso urrun. Proposatzen den metodoan osagai ez-periodikoan hobekuntzak egin nahi dira eta ziur aski emaitza hobeak lortuko dira. Hurrengo lanetan ebaluaketa formalagoak egingo dira, ebaluatzaile gehiagok parte hartuz.

## 5. 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.

## 6. Aipamenak

S. Ahmadi, A.S. Spanias, “Low bit-rate speech coding based on an improved sinusoidal model”, Speech Communication, vol.34, pp.369-390, 2001.

E. Banos, D. Erro, A. Bonafonte, A. Moreno, “Flexible harmonic/stochastic modeling for HMM-based speech synthesis”, Proc. V Jornadas en Tecnologías del Habla, pp.145-148, 2008.

J.P. Cabral, S. Renals, K. Richmond, J. Yamagishi, “Glottal Spectral Separation for Parametric Speech Synthesis”, Proc. Interspeech, pp.1829-1832, 2008.

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.

P. Lanchantin, G. Degottex, X.Rodet, “A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method”, Proc. ICASSP, pp.4630-4633, 2010.

E.R. Banga, C. García-Mateo, X. Fernández-Salgado, “Concatenative Text-to-Speech Synthesis based on Sinusoidal Modeling”, chapter in “Improvements in Speech Synthesis”, John Wiley and Sons, pp.52-63, 2001.

P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound”, Proc. of the Institute of Phonetic Sciences, University of Amsterdam, vol.17, pp.97–110, 1993.

O. Cappé, J. Laroche, E. Moulines, “Regularized estimation of cepstrum envelope from discrete frequency points”, IEEE Workshop on Apps. Signal Proc. to Audio & Acoustics, pp.213-216, 1995.

D. Erro, A. Moreno, A. Bonafonte, “Flexible harmonic/ stochastic speech synthesis”, Proc. 6th ISCA Speech Synthesis Workshop, pp.194-199, 2007.

D. Erro, A. Moreno, A. Bonafonte, “Voice Conversion Based on Weighted Frequency Warping”, IEEE Trans. Audio, Speech, & Language Processing, vol.18 , no.5, pp.922-931, 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.

C. Hemptinne, "Integration of the harmonic plus noise model into the hidden Markov model-based speech synthesis system", Master thesis, IDIAP Research Institute, 2006.

A. Kain, “High Resolution Voice Transformation”, Ph.D. thesis, OGI School of Science & Engineering at Oregon Health & Science University, 2001.

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.

J. Laroche, Y. Stylianou, E. Moulines, "HNM: a simple, efficient harmonic+noise model for speech", IEEE Workshop on Apps. Signal Proc. to Audio & Acoustics, pp.169-172, 1993.

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.

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, 2010

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.

Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis”, IEEE Trans. Audio, Speech, & Language Processing, vol.9, no.1, pp.21-29, 2001.

Y. Stylianou, O. Cappé, E. Moulines, “Continuous Probabilistic Transform for Voice Conversion”, IEEE Trans. Speech & Audio Processing, vol.6, no.2, pp.131-142, 1998.

X. Sun, F. Plante, B.M.G. Cheetham, K.W.T. Wong, “Phase modelling of speech excitation for low bit-rate sinusoidal transform coding”, Proc. ICASSP, vol.3, pp.1691-1694, 1997.

T. Toda, K. Tokuda, “Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM”, Proc. ICASSP, pp.3925–3928, 2008.

T. Toda, A.W. Black, K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory”, IEEE Trans. Audio, Speech & Language Processing, vol.15 , no.8, pp.2222-2235, 2007.

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.

J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, S. Renals, “A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis”, IEEE Trans. Audio, Speech, & Language Processing, vol.17, no.6, pp.1208-1230, 2009.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”, Proc. Eurospeech, pp.2347–2350, 1999.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Mixed excitation for HMM-based speech synthesis”, Proc. Eurospeech, pp.2263–2266, 2001.

S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, “The HTK Book Version 3.4”, Cambridge University Press, 2006.

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.

H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, “The HMM-based speech synthesis system version 2.0”, Proc. 6th ISCA Speech Synthesis Workshop, 2007.