HMMetan oinarritutako euskararako testu-ahots bihurketa, HTS erabiliz

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

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

Abstract

This paper shows how an HMM-based speech synthesizer in Basque language has been built using HTS and AhoTTS (the TTS system developed at Aholab). The resulting system, which is being used only for research purposes at present, has a highly satisfactory performance

Laburpena

Artikulu honek azaltzen du nola garatu den HMMetan oinarritutako euskararako testu-ahots bihurgailua, HTS eta AhoTTS (Aholab taldearen TTS sistema) erabiliz. Garatutako bihurgailua ikerketa-lanetarako baino ez da erabiltzen ari gaur egun, baina oso emaitza onak eman ditu.

Keywords: Text to speech synthesis, HMM synthesis

Hitz gakoak: Testu-ahots bihurketa, HMM sintesia.

1. Sarrera

Blizzard Challenge (Black & Tokuda, 2005) kanpainaren emaitzek erakusten dutenez, Markoven eredu ezkutuetan (HMM, *Hidden Markov Models*) oinarritutako ahots-sintesi sistemak (Zen et al., 2009) gero eta erabiliagoak dira, eta gero eta emaitza hobeak dituzte. Corpus bidezko sintesi-sistemak, aldiz, galtzen ari dira, hein berean, azken urteotan hartu duten garrantzia (Hunt & Black, 1996)(Raux & Black, 2003). 2008an egindako Albayzin ebaluazioak (Sainz et al., 2010) ere gauza bera erakutsi du. HMM sintesi sistemek, tramaka kalkulatutako espektroan eta seinalearen adierazpen parametrikoan oinarrituta, hizlariaren ahotsaren ezaugarri akustikoak modelatzen dituzte, corpus batez entrenatutako testuinguruaren araberako HMM anizkoitzak (*multi-stream context dependent HMM*, CD-HMM) erabiliz. Sintesia egiteko unean, haren testuinguru fonetiko eta prosodikoa kalkulatzen da sarrerako testutik abiatuta, eta HMM eredu bat sortzen da entrenatutako CD-HMM multzotik. Hala, ereduarekiko antz handiena duen parametro-bektore sekuentzia itzultzen du sintesi-sistemak, eta seinale sintetikoa eraikitzen da alderantzizko parametrizazioaren bidez. HMM bidezko sintesi-sistemek duten abantaila nagusia malgutasuna da; izan ere, egokitu egin daitezke entrenatutako ereduak, hartara ahots desberdinak, hitz egiteko era desberdinak, emozio desberdinak eta abar lortzeko.

Sintesi-teknologia honek lortu duen arrakastan erabateko zerikusia izan du *HMM-based Synthesis System* (HTS) (Yoshimura et al., 1999)(Zen et al., 2007) izeneko sistema askatzeak; izan ere, oso onak dira HTS sistemaren emaitzak, azken urteotan oinarrizko sisteman txertatu diren hainbat hobekuntzari esker: f0 modelatzeko banaketa espazioaniztuna erabiltzea (Tokuda et al., 2002), estatikaren eta dinamikaren arteko erlazio esplizituen bidez ibilbideen ereduak sortzea (Zen et al., 2006), egoera esplizituen iraupen-banaketak baliatzea (Zen et al., 2007), parametroak bariantza orokorra kontuan hartuta sortzea (Toda & Tokuda, 2007), *vocoding* teknika sendoak erabiltzea (Zen et al.,2007), eta abar. Emaitza horiei erreparatuta, munduko ikerketa-talde askok garatu dituzte HTSetan oinarritutako sintesi-sistemak, 30 hizkuntza eta dialekto baino gehiagorako ((Zen et al., 2009) artikuluan ageri da zerrenda osoa). Hizkuntza iberiarrei dagokienez, gaztelaniarako (Gonzalvo et al., 2007)(Barra-Chicote et al., 2008), katalanerako (Bonafonte et al., 2009) eta, halaber, portugeserako (Barros et al., 2005) garatu dira HMM bidezko sintesi-sistemak.

Artikulu honek beste hizkuntza bat gehitzen dio zerrenda horri: euskara. Deskribatuko den sisteman, HTS eta AhoTTS (AhoLab taldearen testu-ahots bihurketa-sistema (Hernaez et al., 2001)(Sainz et al., 2008)) sistemak uztartu dira. Hurrengo atalean, sistema horren oinarria zertan datzan azalduko da, eta, ondoren, sistemaren emaitzei buruzko zenbait alderdiri buruz ere jardungo da.

2. AhoTTS-tik AhoHTS-ra

2.1. AhoTTS sistemaren deskribapen laburra

AhoLab laborategian 1997az geroztik garatu den testu-ahots bihurketa-sistema da AhoTTS. Izatez eleanitza bada ere (euskararako (Hernaez et al., 2001), gaztelaniarako (Sainz et al., 2008) nahiz ingeleserako (Sainz et el.,2009) ahotsak garatu dira orain arte), euskararako egin da ahaleginik handiena, eta erreferentziazko sistema bihurtu da, gaur egun, hizkuntza horretan. AhoTTS sistema hiru modulu nagusik osatzen dute: 1) Modulu linguistikoa: testua eta hizkuntza prozesatzeko modulua; 2) Modulu prosodikoa: prosodia aurrez iragartzeko modulua; 3) Seinalea sortzeko modulua. Jarraian, labur-labur deskribatuko ditugu hiru moduluok.

Modulu linguistikoak sarrerako testua irakurri eta dagokion fonema segida sortzen du. Gainera, maila desberdinetako informazio linguistikoa ere gehitzen dio. Modulu horrek honako zeregin hauek betetzen ditu: testua normalizatzea, esaldiak mugatzea, kategoria gramatikala (POS, *part-of-speech*) esleitzea, silabak mugatzea, azentua ezartzea eta fonemak zehaztea.

Modulu prosodikoak, bigarrenak alegia, lehen moduluak sortzen duen informazio fonetiko eta linguistikoa erabiliz, ingerada prosodiko bat sortzen du, hiru ezaugarri kontuan izanda: intonazioa, iraupena eta energia. Intonazioaren balioak kalkulatzeko, hiru estrategia garatu dira: haran-gailur eredua, sinpleena; zuhaitzetan eta Fujisakiren kurbetan oinarritutako eredua (Navas et al.,2002), pittin bat sofistikatuagoa; eta corpus bidez ingurua hautatzeko eredua (Raux & Black, 2003). Iraupenen balioak kalkulatzeko, sailkapen- eta erregresio-zuhaitzak (*Classification and Regression Trees*, CART) erabiltzen dira.

Seinalea sortzeko moduluak –hirugarrenak– aurreko bi moduluek sortutako informazioa jaso eta seinale sintetikoa sortzen du. AhoTTSren gaur egungo bertsioak unitateak hautatzeko teknika erabiltzen du (Hunt & Black, 1996).

AhoTTS sisteman sintesi parametrikoa txertatu ahal izateko, HTSa erabili behar da bigarren eta hirugarren moduluen ordez. HTSak, entrenatutako ereduak erabiliz, seinalearen prosodia eta espektroa sor ditzake, baina ez da gai analisi linguistikoa egiteko. Hortaz, AhoTTS sistemak sortzen ditu HTSak behar dituen etiketa linguistiko guztiak. Dena dela, itzuli egin behar da AhoTTSko lehen moduluaren irteera, etiketa linguistikoek formatu egokia izan dezaten.

2.2. Euskarari buruzko iruzkin batzuk

Euskaraz, beste hizkuntza askotan bezala, maila askotan dago kokatuta informazio linguistikoa: fonemetan, silabetan, hitzetan, azentu-taldeetan, esaldietan eta abarretan. Azentu-taldea silaba azentudun bakar baten inguruan ahoskatzen diren silabek osatzen dute (Möbius et al., 1993); zenbait hizkuntzatan, bereziki hizkuntza iberiarretan (Escudero, 2002)(Navas, 2003)(Agüero et al., 2006)(Campillo & Banga, 2005), hitzek osatzen dute azentu-taldea, eta haietariko batek izan ohi du azentua. Euskara hizkuntza malgukaria eta eranskaria izanik (Hualde & Ortiz de Urbina, 2003), azentu taldea askotan hitz bakar batek osatzen du; oso gutxitan gertatzen da azentu-taldean hurrengo hitzak ere sartzea; esaterako, aditz-laguntzaile laburren kasuan, erakusleen kasuan eta zenbait zenbakiren kasuan. Hortaz, baliteke maila linguistiko desberdinen informazioa erredundantea izatea, baina horrek ez du batere kalterik eragiten sisteman; izan ere, HTSak informaziorik garrantzitsuena solik hautatzen du CD-HMM-ak entrenatzeko. Azentu-taldeak kontuan izateak badu abantaila bat: erraztu egiten du gainerako hizkuntza iberiarrak –gaztelera, adibidez– sistema honetan sartzeko aukera.

Euskara hizkuntza malgukaria eta eranskaria izatearen beste ondorio bat da azentu bat baino gehiagoko hitz luzeak agertzen direla. Hitz horiek ondo maneiatzea oso zaila da; gainera, handia da euskarazko dialekto kopurua (zazpi euskalki nagusi eta 50 barietate baino gehiago daude), eta horrek are handiagotzen du intonazioaren aldakortasuna. Horrenbestez, lan honetan ez dira kontuan hartu azentu anizkoitzak; jo da sistemak bigarren mailako azentua datu akustikoetatik eta etiketetatik ikasten duela.

2.3. Testuinguru-etiketak sortzea, AhoTTS erabiliz

AhoTTS sistemak sortzen dituen ezaugarri linguistikoen artean, honako hauek kodetu dira HTS sistemak behar dituen testuinguru-etiketetarako:

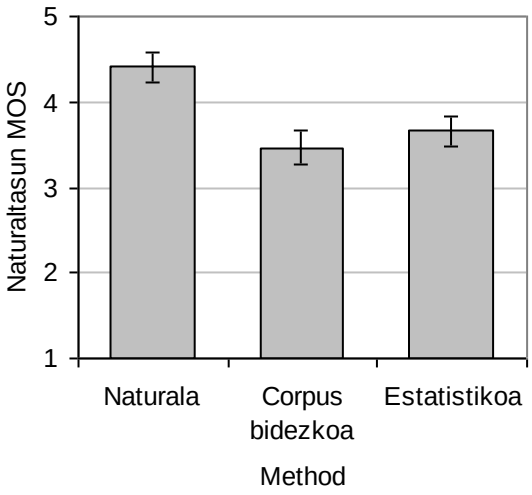
- Fonema-mailan:
    - Uneko fonemaren SAMPA etiketa.
    - Uneko fonematik abiatuta, 2 ezkerrerago eta 2 eskuinerago dauden fonemen SAMPA etiketak.
    - Uneko fonemak uneko silaban duen kokalekua (bai hasieratik, bai amaieratik neurtua).
    - Uneko fonemaren kokalekua, bai aurreko etenalditik, bai hurrengo etenaldira.
- Silaba-mailan:
    - Uneko, aurreko eta hurrengo silaben fonema kopurua.
    - Azentua uneko, aurreko eta hurrengo silabetan.
    - Indargunea uneko, aurreko eta hurrengo silabetan.

- Uneko silabaren kokalekua uneko hitzean (bai hasieratik, bai amaieratik neurtuta).
- Uneko silabaren kokalekua uneko azentu-taldean.
- Uneko silabaren kokalekua uneko esaldian.
- Uneko silabaren kokalekua, bai aurreko etenaldidik, bai hurrengo etenaldira.

- Hitz-mailan:
  - Uneko, aurreko eta hurrengo hitzen kategoria gramatikalen (POS, *part-of-speech*) etiketa sinplifikatua (content/function).
  - Uneko, aurreko eta hurrengo hitzen silaba kopurua.
  - Uneko hitzaren kokalekua esaldian (hasieratik eta amaieratik neurtua).
  - Uneko hitzaren kokalekua, bai aurreko etenetik, bai hurrengo etenera.
- Azentu-taldearen mailan:
  - Uneko, aurreko eta hurrengo azentu-taldeen mota, azentuaren kokalekuaren arabera.
  - Uneko, aurreko eta hurrengo azentu-taldeen silaba kopurua.
  - Azentu-taldearen kokalekua esaldian (bai hasieratik, bai amaieratik neurtua).
  - Uneko azentu-taldearen kokalekua, bai aurreko etenaldidik, bai hurrengo etenaldira.
- Eten-testuinguruaren mailan:
  - Aurreko eta hurrengo eten mota.
  - Ezkerreranzko eta eskuineranzko eten kopurua.
- Esaldi-mailan:
  - Esaldi mota.
  - Fonema kopurua.
  - Silaba kopurua.
  - Hitz kopurua.
  - Azentu-talde kopurua.
  - Eten kopurua.
  - Esaldiaren emozioa.

3. Emaitzak

Sistemaren emaitzak ebaluatzeko, *Mean Opinion Score* (MOS) probak erabili ziren, emakumezko batek euskaraz eta estilo neutroan irakurritako 2.000 esaldi laburrez osatutako datu-base batez (2 orduko iraupenekoa, gutxi gorabehera). Hamazortzi entzule boluntariok hartu zuten parte ebaluazioan (horietariko seik bazuten eskarmentua testu-ahots bihurgailuen alorrean), eta, hamarna esaldi sintetiko entzunda, 1-5 arteko puntu-eskala batez ebaluatu zituzten: 1 puntuaren esanahia "oso naturaltasun gutxikoa" eta 5 puntuarena "oso naturaltasun handikoa" izanik. Ahotsetik f0 parametroa, 40 MFCC koefiziente eta banden ez-periodikotasunen 5 koefiziente sortzeko, STRAIGHT bidezko *vocoder*ra erabili zen, eta datu horiekin entrenatu ziren ereduak. Azken HTS bertsioak[1] gomendatzen duen moduan, CD-HMMen bidez modelatu ziren inguratzaile espektrala eta osagai ez-periodikoa (baita haien lehen eta bigarren deribatuak ere); bestalde, f0 parametroa espazio anitzeko probabilitate-banaketen bidez modelatu zen. Artikuluan deskribatutako sistemarekin batera, ahots naturala eta corpus bidezko sistema ere ebaluatu ziren, bata eta besteak alderatzeko.

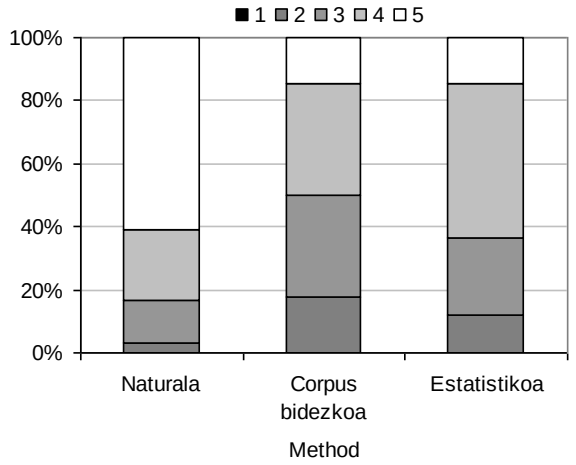


1. Irudia Sistema bakoitzaren naturaltasun-MOSaren puntuazioa, %95eko konfiantza-tartearekin

Artikulu honetan deskribatutako sistemak lortu duen naturaltasun-MOSa 3,7 puntu da, corpus bidezkoak lortu duena baino pixka bat handiagoa, bataren zein bestearen konfiantza-tarteak gainjarrita egonik ere (ikus 1. irudia). Zenbait esaldi solte hartu eta azterketa sakonago bat eginik, ondorioztatu da

[1] [Online], "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/

ebaluatzaileek nahiago dutela sistema estatistikoa uztartze-zarata dagoenean. Fenomeno bera ikusi zen 2008ko Albayzin kanpainan ere (Sainz et al., 2010). 2. irudian puntuazioen banaketa ageri da: sistema estatistikoak kasuen %12an bakarrik lortu zuen 3 puntu baino gutxiago. Horrek adierazten du HTS sistema gai dela ahots sintetikoa era atsegin eta egonkorrean sortzeko, beste hizkuntza askotan gertatzen den bezala (Zen et al., 2009). Corpus bidezko sistemak sistema estatistikoak baino kasu gehiagotan lortu zuen 5 puntu; datu hori ere aurreko ikerketetan agertu zen, bai eta Blizzard Challenge ebaluazioetan ere.



2. Irudia Naturaltasun-MOS probako puntuazioaren banaketa

1. taulak erakusten du zer heinetan dauden erlazionatuta testuinguru-menpekotasuna atzematen duten zuhaitzen nodoak eta testuinguru-maila bakoitza. Taulan ageri denez, fonemek eta silabek dute informazio adierazgarri gehien. Hitz- eta azentu-taldeei dagokienez, biek daukate eragin handia prosodian; hala ere, azentu-taldeak hitzak baino lehenago agertzen direnez, ondoriozta daiteke azentu-taldeek hitzek baino informazio adierazgarriagoa dutela euskaraz.

	Espektroa	*Pitch*a	Iraupena
Fonema	93.87 (0)	43.56 (0)	72.94 (0)
Silaba	3.54 (3)	24.97 (1)	12.94 (2)
Hitza	0.38 (5)	9.14 (3)	4.51 (3)
Azentu-taldea	0.63 (4)	12.29 (0)	5.49 (0)
Eten-testuingurua	0.94 (2)	2.58 (2)	0.98 (1)
Esaldia	0.64 (4)	7.47 (2)	3.14 (4)

1. Taula: Testuinguru-etiketen maila bakoitzeko zuhaitz-nodoak. Parentesi artean: adarkatzearen lehen maila, adarkatzerik badenean.

4. Ondorioak

Euskararako HMM bidezko sintesi-sistema bat garatu da, HTS sistema eta AhoTTSko modulu linguistikoa erabiliz. Sortutako ahotsak sintesi estatistikoaren ohiko mugak dituen arren, ebaluatzaileek emandako puntuek azaltzen dute haien ustez nahiko naturala dela.

Gaur egun, ikerketa-taldeak *vocoder* teknika berriak aztertzen ari dira, ahots sintetikoaren kalitatea hobetzeko.

5. Esker onak

Lan hau UPV/EHUren laguntzaz (doktoreak espezializatzeko laguntza), Espainiako Zientzia eta Berrikuntza Ministerioko laguntzaz (*Buceador* proiektua, TEC2009-14094-C04-02) eta Eusko Jaurlaritzaren laguntzaz (*Berbatek*, IE09-262) egin da.

6. Aipamenak

Agüero, P.D. ; J. Adell, A. Bonafonte, “Prosody Generation for Speech-to-Speech Translation”, Proc. ICASSP, pp.557-560, 2006.

Barra-Chicote, R.; J. Yamagishi, J.M. Montero, S. King, S. Lufti, J. Macías-Guarasa, “Generación de una voz sintética en castellano basada en HSMM para la evaluación Albayzin 2008: conversión texto a voz”, Proc. V Jornadas en Tecnología del Habla, pp.115-118, 2008.

Barros, M.; R. Maia, K. Tokuda, D. Freitas, F. Resende Jr., “HMM-based European Portuguese speech synthesis”, Proc. Interspeech, pp.2581–2584, 2005.

Black, A.W.; K. Tokuda, "The Blizzard Challenge – 2005: evaluating corpus-based speech synthesis n common datasets", Proc. Interspeech, pp.77-80, 2005.

Bonafonte, A.; L. Aguilar, I. Esquerra, S. Oller, A. Moreno, “Recent work on the FESTCAT database for speech synthesis”, I Joint SIG-IL / Microsoft Workshop on Speech and Language Technologies for Iberian Languages, 2009.

Campillo, F.; E.R. Banga, “A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems”, Speech Communication, vol.48, pp.941-956, 2005.

Escudero, D. “Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversión Texto-Voz en Español”, PhD thesis, Universidad de Valladolid, 2002.

Gonzalvo, X.; 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.

Hernaez, I.; E. Navas, J.L. Murugarren, B. Etxebarria, “Description of the AhoTTS system for the Basque language”, Proc. 4th ISCA Speech Synthesis Workshop, 2001.

Hualde, J.I.; J. Ortiz De Urbina (Eds.), “A Grammar of Basque”, Mouton de Gruyter, Berlin, 2003.

Hunt, A.; A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, Proc. ICASSP, vol. 1 pp. 373-376, 1996.

Möbius, B.; M. Pätzold, W. Hess. “Analysis and synthesis of German F0 contours by means of Fujisaki’s model”. Speech Communication, vol.13, pp. 53-61, 1993.

Navas, E.; I. Hernaez, J. Sanchez, “Subjective evaluation of synthetic intonation”, Proc. IEEE Workshop on Speech Synthesis, pp.23-26, 2002.

Navas, E.; I. Hernáez, J. Sánchez, “Predicting Segmental Durations for Basque Using CARTs”, Proc. 15th International Congress of Phonetic Sciences, pp.2083-2086, 2003.

Navas, E. “Standard Basque Prosodic Modeling for Text to Speech Conversion”, PhD thesis, University of the Basque Country, 2003.

Raux, A.; A. Black, “A unit selection approach to F0 modeling and its application to emphasis”, Proc. ASRU, pp. 700- 705, 2003.

Sainz, I.; E. Navas, I. Hernáez, A. Bonafonte, F. Campillo, “TTS Evaluation Campaign with a Common Spanish Database”, Proc. 7th International Language Resources and Evaluation Conference, pp. 2155-2160, 2010.

Sainz, I.; I. Hernáez, E. Navas, J. Sanchez, I. Luengo, I. Saratxaga, I. Odriozola, E. de Bilbao, D. Erro, “Descripción del Conversor de Texto a Voz AhoTTS Presentado a la Evaluación Albayzin TTS 2008”, Proc. V Jornadas en Tecnología del Habla, pp.96-99, 2008.

Sainz, I.; D. Erro, E. Navas, I. Hernáez, I. Saratxaga, I. Luengo, I. Odriozola, “The AHOLAB Blizzard Challenge 2009 Entry”, Blizzard Challenge 2009 Workshop, 2009.

Toda, T.; K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis”, IEICE Trans. Inf. Syst. E90-D (5), pp.816–824, 2007.

Tokuda, K.; T. Masuko, N. Miyazaki, T. Kobayashi, “Multi-space probability distribution HMM”, IEICE Trans. Inf. Syst. E85-D (3), pp.455–464, 2002.

Yoshimura, T.; 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.

Zen, H.; K. Tokuda, A.W. Black, “Statistical parametric speech synthesis”, Speech Communication, vol.51, no.11, pp.1039-1064, 2009.

Zen, H.; 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.

Zen, H.; K. Tokuda, T. Kitamura, “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences”, Computer, Speech and Language, vol.21(1), pp.153–173, 2006.

Zen, H.; K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “A hidden semi-Markov model-based speech synthesis system”, IEICE Trans. Inf. Syst. E90-D (5), pp.825–834, 2007.

Zen, H.; 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.