TTS ikuspegi hibrido bat modulu akustiko eta prosodikoentzat

Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernáez

AHOLAB Signal Processing Laboratory
{inaki, derro, eva, inma}@aholab.ehu.es

Abstract

Unit selection (US) TTSs generate quite natural speech but highly variable in quality. Statistical parametric (SP) systems offer far more consistent quality but reduced naturalness due to its vocoding nature. We present a hybrid approach (HA) that tries to improve the overall naturalness combining both synthesis methods. Contrary to other works, the fusion of methods is performed both in prosody and acoustic modules yielding a more robust prosody prediction and achieving greater naturalness. Objective and subjective experiments show the validity of our procedure.

Laburpena

Unitate aukeraketako (UA) TTSek (Testu Ahots Bihurgailuek) ahots nahiko naturala sortzen dute, baina kalitate oso aldakorrarekin. Sistema estatistiko-parametrikoek (EP), berriz, kalitate askoz trinkoagoa eskaintzen dute, baina murriztutako naturaltasunarekin beren vocoderraren izaera dela eta. Bi sintesi-metodoak konbinatuz naturaltasun orokorra hobetzen saiatzen den ikuspegi hibridoa (IH) aurkezten dugu hemen. Beste lan batzuetan ez bezala, metodoen bat-egitea bai modulu prosodikoan bai modulu akustikoan gauzatzen da, eta prosodiaren iragarpen sendoagoa eta naturaltasun handiagoa lortzen da. Esperimentu objektibo zein subjektiboek gure prozeduraren baliagarritasuna erakusten dute.

Keywords: speech synthesis, unit selection, statistical synthesis, hybrid system

Gako hitzak: ahots sintesia, unitate aukeraketa, sintesi estatistikoa, sintema hibridoa

1. Sarrera

90ko hamarkadaren erdialdetik aurrera, TTS-en garapenean UA-n oinarritutako ikuspegi kateatua teknika nagusia izan da. "Unitate-aukeraketaren oinarrizko ideia zera da: ahots naturaleko datu-base batetik hitzak baino txikiagoak diren unitate egokiak aukeratuz, soinu naturala duten esaldi berriak sintetizatu ditzakegula." (Black, 2002). (Hunt eta Black, 1996)k datu-basearen unitate hautagaien segida aurkitzeko Viterbi bilaketa generikoa proposatzen du, non bi azpikostuz osatutako kostu funtzio bat minimizatzen den: kostu objektiboak eta kateatzeko kostuak. Kostu objektiboak unitate hautagaiek eta behar den unitateak ea zein ondo bat etor daitezkeen balioztatzen du. Gehienetan, kostu objektiboa neurtzeko testuinguru linguistiko eta prosodikoak erabiltzen dira, azpikostuen edo aurre-multzokatutako erabaki-zuhaitzen bitartez (Donovan eta Woodland, 1995) (Black eta Taylor, 1997). Kateatze-kostuak bi unitateren arteko bat-egitearen bikaintasuna neurtzen du, normalean distantzia espektralen, energiazko distantzien eta tonuaren distantziaren arabera. Tamaina desberdineko unitateak proposatu dira, baina orokortuz, zenbat eta unitatea luzeagoa izan, gero eta corpus handiagoa behar da (edo aplikazio-eremua txikiagoa). Ikuspegi honen abantaila nagusiena unitateen naturaltasun segmentala babestea da, eremu murriztuetan errendimendu nabarmena lortuz. Hala ere, hainbat desabantaila zerrendatu daitezke: kalitate aldakorra (kateatze txar bakar batek esaldi guztia hondatu dezake), espazio eskakizun altuak eta kostu altuak ahots berri baten sorkuntzan.

EP ahots sintesia adituen esker ona jasotzen ari da mendea aldatu zenetik. Bere premisa akustikoki antzekoak diren segmentuekin egindako batez besteko ereduetatik ahotsa sintetizatzea da. Entrenamendu-fasean, ahots naturaletik parametro espektralak eta kitzikapenezkoak erauzten dira, eta horretaz gain hizkuntza ezaugarriak sortzen dira testuinguruaren mendeko HMMak entrenatzeko. Sintesian zehar, sarrerako testutik inferitutako testuinguru linguistikoak erabiltzen dira eredu egokiak aukeratzeko eta ahots parametroak sortzeko, zeintzuk ahots bihurtuko diren vocoder baten laguntzaz. (Wu eta Wang, 2006)k sorkuntza-akatsa txikiagotzen duen entrenamendu metodoa proposatzen du, zeinen helburua sintesiaren kalitatea hobetzea den. Bai modelatze estatistikoan bai ahotsaren parametrizazioan azken hamarkadan egindako hobekuntzen laburpen zehatza irakurtzeko, begiratu (Zen eta beste, 2009). Ikuspegi honen abantaila nagusiak honakoak hauek dira: trinkotasuna, malgutasuna eta behar duen espazio txikia. Ahots leun eta egonkorra sortzen du, orokortze-ezaugarri onekin. Modelatze estatistikoak ahots eta estilo transformazio errazak posible egiten ditu teknika desberdinen bitartez (egokitzapena, interpolazioa, erregresio aniztuna eta auto-ahotsa). Desabantailak hurrengoak dira: vocoder kalitatea (gutxitutako naturaltasuna, burrunbak) eta gehiegizko leunketa estatistikoa.

Hainbat ahalegin egin dira bi tekniken puntu sendoak TTS hibrido batean konbinatzeko, gehienak

kateatze ikuspegian oinarritutakoak. Lan batzuk (Kawai eta beste, 2004) (Yang eta beste, 2006) (Krstulovic eta beste, 2008.) prosodiaren sorkuntzarako modelatze estatistikoa aukeratu dute, gero modulu akustikoa hornitzeko. (Hirai eta beste, 2007)n HTS (HTS Webgunea, 2012) parametro akustikoak sortzeko erabiltzen da, zeintzuk trama objektibo bezala erabiliko diren distantzia Euklidestarrean oinarritutako UA prozesuan. (Rouibia ta beste, 2005)en hautagaizko difonema antzekoenak aukeratzeko parametro espektralak erabiltzen dituzte soilik. Kostu konputazionala txikiagotzeko, (Hirai eta beste, 2007) (Yan eta beste, 2010)ek Kullback-Leibler dibergentzia erabiltze dute HMM eredu objektibo eta hautagaien artean, unitate aurretiko-aukeraketa fasean.

Beste lan batzuk saiatu dira teknika biak uhin-formaren sorkuntzan konbinatzen. Honela, segmentu naturalen kalitatea mantentzen ahalegintzen dira, baina datu-basean hautagai egokiak aurkitzen ez direnean ahots modelatua erabiltzen dute. Hala ere, trantsizio-zonetan ahots kalitatearen aldaketa entzungarriak badaude, emaitza are okerragoa izan daiteke (Aylett eta Pidcock, 2009). (Breen eta Pollet, 2008)n programazio dinamikoko fase batek erabakitzen du zein den unitate naturalen eta estatistikoki sortutakoen segidarik honena. Konbinazioaren soinu-akatsak txikiagotzeko asmoz, parametro akustikoak birsortzen dira beraien bariantza unitate naturalen segida honenarekin doituz. (Silén eta beste, 2010)en, unitate desegokien aukeraketa saihesteko Viterbiren algoritmo sendoa (Siu eta Chan, 2006) darabilte. Unitate desegokiak modelatutako ahotsarekin ordezkatzen dira orduan. Kalitatearen aldaketak txikiagotzeko asmoz, bai segmentu naturalak bai modelatuak berreraikitzen dira vocoder baten bitartez. (Gonzalvo eta beste, 2009)n HMMetan oinarritutako sintesiaren kalitatea hobetzeko proposamen ezberdina aurkezten da. Lehenik eta behin, HTSrekin entrenatutako zuhaitzetako orrietako unitate natural erabilgarrien artean unitate aukeraketa bat burutzen da egoera mailan. Gero, unitate naturalen media eta bariantza erabiltzen dira modelatutako parametro akustikoen birsorkuntzan, sorkuntza lokaleko akatsa txikiagotuz.

Artikulu honetan kateatze ikuspegian oinarritutako TTS hibridoa aurkezten dugu. Aurretik aipatutako lanak ez bezala, gure arkitekturak UA eta EP sintesia konbinatzen du bai modulu prosodikoan bai akustikoan. 2. sailean, gure oinarri-sistemaren eta HTSan oinarritutakoaren deskribapen laburra egiten da. 3. sailak bi tekniken konbinazioa azaltzen du TTS hibridoaren eraikuntzan. Ebaluazio subjektibo eta objektiboen emaitzak 4. sailean erakusten dira. Eta azkenik, zenbait ondorio aipatzen dira.

2. Aholaben TTS sistema

AhoTTS (Hernaez eta beste, 2001) Aholab laborategia 1995-etik garatzen ari den sintesirako plataforma da, ikerkuntzarako eta merkataritzarako zuzenduta dagoena. Arkitektura modularra dauka, eta C/C++-en idatzita bai UNIX bai Windows sistema eragileetan guztiz erabilgarria da. Orain arte euskararako, gaztelaniarako eta ingeleserako ahots sintetikoak garatu dira. Ondoren, gure oinarri-sistemaren eta HTSn oinarritutako sistemaren ezaugarri nagusiak deskribatzen ditugu.

2.1. Unitate aukeraketako oinarri-sistema

Modulu independiente anizkunetan datza. Lehenak hizkuntzaren menpeko hainbat ataza burutzen ditu: Testuaren normalizazioa, POS etiketaketa, silabifikazioa eta letra-fonema konbertsioa.

Modulu prosodikoak zenbait ataza sekuentzial burutzen ditu. Fonemen iraupenaren aurresanean zscore ereduak entrenatzen dira: Random Forest (RF) (Breiman, 2001) bokaletarako eta CARTs kontsonanteetarako. Aholaben UA intonazio modelatzeak fonema ahotsdunak erabiltzen ditu oinarrizko unitate bezala, (Raux eta Black, 2003)ren antzeko ikuspegia jarraituz eta silaba barruko kateatzeak murriztuz. (Sainz eta beste, 2009) kontsulta ezazu azpimodulu honen deskribapen zehatzagorako. Corpusak haustura intonatiboak (IH) (Campillo eta beste, 2009) markaturik badauzka bere iragarpenerako CART bat entrenatzen da hurrengo ezaugarri sinpleak erabiliz: POS etiketa hiru hitzetako leiho batean, silaba kopurua, azentudun silaba kopurua, hurrengo eta aurreko etenerarte (IH edo isilunea) dauden hitz kopurua. Orduan, IH informazioa iraupenaren eta intonazioaren aurresanean erabiltzen da fonema, silaba eta hitz mailan, iragarpenaren zehaztasuna hobetuz, 4.1 sailean erakusten den moduan.

Irakurri (Sainz eta beste, 2009) motore akustikoan erabilitako ezaugarrien eta azpikostuen diseinuaren azterketa zehatzerako. UA ondoren, aldaketa prosodiko txikiak baino ez dira egiten, pitch-sinkronoak diren gainezarpen eta batuketa tekniken bitartez.

2.2. AhoHTS: HTSn oinarritutako sistema

HTSk ez duenez inongo analisi linguistikorik egiten, AhoTTSren lehenengo moduluaren irteera etiketa formatu egokira bihurtu zen. Testuinguruko etiketetan erabilitako ezaugarrien zerrenda zehatza ikusteko, irakurri (Erro eta beste, 2010) . Trama bakoitzeko bai espektroaren bai kitzikapenaren irudikapen parametrikoa lortzeko, HNMn (Harmonics plus Noise Model) oinarritutako vocoder bat erabiltzen da (Erro eta beste, 2011), vocoder honek ahotsa berreraikitzeko aukera ere ematen baitu.

3. TTS hibridoa

Sistema hibridoaren arkitektura 1. irudian erakusten da. Laburbilduz, HTSren irteera aurresate objektiboa bezala erabiltzen da UA moduluan. HTSren intonazio eta iraupen iragarpenak AhoTTSren modulu prosodikoekin konbinatzen dira, eta parametro espektralak objektiboa eta hautagaizko unitateen arteko distantzia kalkulatzeko erabiltzen dira. Ikuspegi hibrido

hau ereduen sendotasuna eta ahots unitate naturalen kalitate segmentala konbinatzen saiatzen da.

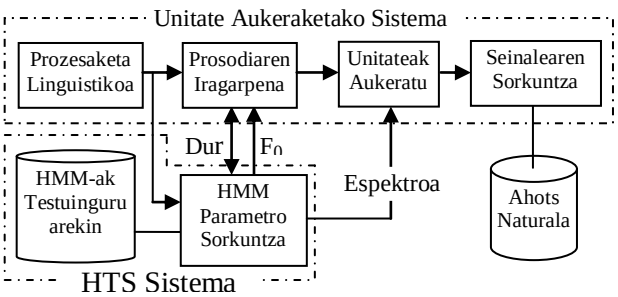
3.1. Modulu prosodikoa

Iragarpen prosodikoko moduluari dagokionez, IH gehienak bakarrik HMMen aurresateaz fidatzen dira. Hala ere, iraupen-aurresan hobeak lor daitezke teknika desberdinen fusioaren bidez (Lazaridis eta beste, 2011). Gainera, (Hirai eta beste, 2007)n MOS (Mean Opinion Score) onena lortzen dute HMMetan oinarritutako intonazio kurbari kanpoko iraupena inposatuz.

Aholaben IHan, HTSko eta CART/RFko iraupen aurresanaren konbinazio lineala egiten da. Hobekuntza erakusten duten neurri objektiboek 4.1.1 sailean aurkezten dira. Nabarmendu behar da iraupenaren fusioa noranzko bikoa dela (hau da, irteerak bai modulu prosodiko estandarra bai HMM parametro sortzailea elikatzen ditu). Lehenik eta behin, fonemen iraupenak HTS motore barruan iragartzen dira. Orduan, iragarpena linealki konbinatzen da modulu prosodiko estandarrarenarekin, fonema mailan inposatuz. Azkenik, HTSk egoera bakoitzaren luzera aurresaten du jada lehenetsita zegoen fonemaren luzera barruan. Operazio honekin lortzen duguna denborazko sinkronizazioa erraztea da, modulu akustikoan egiten den ondorengo intonazio eta espektro konparaziorako. Prozedura hau jarraituz gero, entzute-test informalek HTSn oinarritutako sistemaren naturaltasunaren hobekuntza ere erakusten dute.

Bi intonazio kurben fusioa hainbat fasetan banatzen da. Lehenik eta behin f0 balioak eremu ahoskabeetan interpolatzen dira eta bi kurbak fonema mailan segmentatzen dira, bakarrik kanonikoki ahostun diren fonemen f_0 balioak gordez. Fonema ahostun bakoitzeko 3 puntuko pitch estilizazioa egiten da. Azkenik, aurrez lerrokatutako fonema luzerako pitch zatien arteko haztatutako konbinazio lineala egiten da. Metodo sinple honek hobekuntza txikiak lortzen ditu 4.1.2 sailean erakusten diren neurketa objektiboetan, eta estatistikoki adierazgarriak direnak 4.2 saileko kutxa-beltz test subjektiboan.

Goian azaldutako fusio prosodikoko prozesuan, konbinazio linealaren pisuak eskuz doitu ziren, HTSko pitch-iragarpenari eta CART/RF iraupen-iragarpenari garrantzi handiagoa emanez hurrenez hurren, 4.1 sailean aurkeztutako test objektiboen emaitzen arabera.



1. Irudia: TTS Hibridoaren Arkitektura.

3.2. Modulu akustikoa

UA prozesuan, TTS hibrido gehienak bakarrik EP sistemak sortutako ibilbide akustikoaz fidatzen dira. Aipaturiko aukeran ez bezala, ohiko azpikostu objektiboak mantentzen ditugu (linguistikoak eta prosodikoak) eta azpikostu berri bat gehitzen dugu:

Distantzia Espektrala: Tramaz trama kalkulatutako distantzia Euklidestarra da, objektiboa (HTS irteera) eta hautagai unitateen artean DTW (Rouibia eta beste, 2005) lerrokatzearen ondoren. Distantzia eskuz haztatzen da hiru klase fonetiko murriztuen arabera: bokalak, kontsonante ahostunak eta kontsonante ahoskabeak.

Ikuspegi honen abantaila nagusia zera da: unitateak aukeratzerako orduan beraien antzekotasun akustikoa esplizituki (HTS) eta inplizituki modelatuz gero, (azpikostu objektiboa) prozedura sendoagoa dirudiela. Distantzia espektralaren ekarpen nagusietako bat "unitate txarrak" (txarto etiketatuak edo ahoskatuak) aukeratzea saihestea da, sintesi sendoagoa lortzen da. Nola distantzia espektralen kalkulua bereziki neketsua den, aurretiko aukeraketa-fasean bakarrik azpikostu linguistiko eta prosodikoak (eta beraz askoz ere sinpleagoak) erabiltzen dira, horrela sintesi prozesua bizkortuz.

4. Ebaluaketak

Sistema hibrido berriaren errendimendua egiaztatzeko, gaztelaniazko ahots bat garatu zen. Corpusa *Albayzin 2010 TTS* ebaluaketaren (Méndez eta beste, 2010) antolatzaileek hornitua izan zen. Gizonezko ahots baten bi orduko grabazioz osatuta zegoen, estilo neutroan eta 16 kHz-etan. Material osagarriaren barruan segmentazio automatikoa eta IH etiketak zeuden. Segmentazio markak automatikoki bikaindu eta gero, Aholaben UA TTSaren ohiko ahots garapen prozesua burutu genuen. EP ahotsari dagokionez, HTS demoaren script-ak erabili aurretik, jadanik 2.2 sailean aipatutako vocoderrarekin parametrizatu ziren seinaleak (40 MFCC + f0 lortuz).

Sistema hibridoaren kalitatea balioztatzeko ebaluaketa objektiboak zein subjektiboak burutu ziren.

4.1. Ebaluaketa objektiboak

Albayzin 2010 TTS ebaluaketaren antolatzaileek testean erabilitako 350 esaldien grabazio naturalak banatu zituzten kanpaina bukatu zenean. Azpimarratu behar da esaldi horiek ez zirela erabili ahotsaren garapenean. Datu hauek automatikoki segmentatu ondoren intonazio kurbak lortu genituen pitch detekziorako hiru algoritmo ezberdin konbinatuz (pthcdp (Luengo eta beste, 2007), praat eta get_f0). Orduan, prosodia naturala eta sintetikoa konparatu genuen zenbait iragarpen metodorako. Hiru neurri erabili ziren erreferentzia naturala eta aurresandako prosodiaren arteko desadostasuna neurtzeko: Batez besteko Errore Koadratikoa (BEK), Batez besteko

Errore Absolutua (BEA) eta Pearson Korrelazio koefizientea. BEK eremutik kanpoko balioekin sentikorragoa izanik, gertatutako errore larriak balioztatzeko erabilgarria da.

4.1.1. Iraupenaren iragarpena

1. taulan lau iragarpen metodoen erroreak erakusgarri daude. Wilcoxon signed rank test-arekin (α=0.05) frogatu genuen metodo guztien arteko ezberdintasunak estatistikoki adierazgarriak zirela, hiru neurrientzat. HTSren iraupenaren iragarpenak errendimendurik txarrena dauka baina bere konbinazioak metodo hibridoaren barnean CART+RFren iragarpena hobetzen du. Gainera, IH informazioak badirudi metodo guztien eraginkortasuna handitzen duela (nahiz eta taulan metodo hibridoaren kasurako bakarrik erakusten den). IH fenomeno garrantzitsua da fonemen iraupenarekin (ad. fonemen luzatzea), intonazio kurbarekin (ad. f0 berrezartzea) edo ezaugarri espektralekin (ad. ahosdura erlaxatua) erlazionatuta dagoena. Eta nahiz eta bere iragarpenean sartze faltsuak egon daitezkeen, fenomeno fina denez (behintzat isiluneen sartzearekin konparatuz gero), ez da sumatzen horrelako akatsek prosodiaren iragarpenean eragin larria dutenik.estu normala hemen.

	BEK (ms)	BEA (ms)	Pearson
CART+RF IHrekin	18,09	11,78	0,767
HTS IHrekin	20.78	14.28	0.681
Hibridoa IHrekin	17.95	11.84	0.768
Hibridoa IBrik gabe	18.37	12.16	0.720

1. taula. Fonemen iraupenaren iragarpena hainbat metodo erabiliz, erreferentzia naturalarekin konparatuz.

Kurba sintetikoak eta naturalak konparatu aurretik, denbora-lerrokatze bat egiten da fonema ahostunen mailan, eta pitch balioak milisegundo bakoitzerako lortzen dira. Iraupenean erabilitako hiru neurri berberak ikusi daitezke 2. taulan.

Nahiz eta metodo hibridoak HTSren iragarpena pixka bat hobetzen duen, Wilcoxon sign rank testaren arabera, hobekuntzak ez dira estatistikoki adierazgarriak (baina beste metodo guztien arteko ezberdintasunak adierazgarriak dira). Berriro ere, IH informazioak emaitzak hobetzen ditu eta IHri dagokien ezaugarriek paper nahiko garrantzitsua izaten dute HTSko f_0 zuhaitzetan. Gure UA intonazio moduluak emaitzarik txarrenak aurkezten ditu neurri objektiboetan, baina UA moduluarekin batera lan egiteko diseinaturik dagoenez corpusean dauden testuinguru espezifikoko fonemen pitch-eremu hurbila aukeratzen du normalean. Aholaben modulu akustikoak naturaltasun segmentala mantentzeko bakarrik aldaketa prosodiko txikiak burutzen dituela kontuan hartuz, kutxa beltzeko ebaluaketa subjektibo gehiago prestatu genituen HTSren eta hibridoaren prosodiaren eragina sintetizatutako seinalean konparatzeko asmoz. Emaitzak 4.2. sailean aurkezten dira.

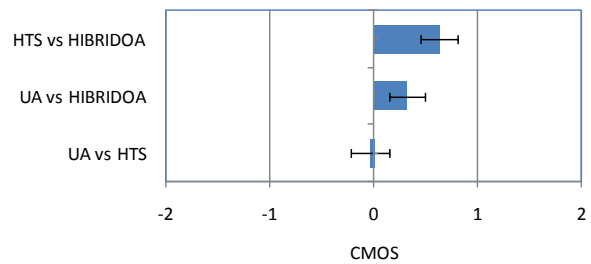
	BEK (Hz)	BEA (Hz)	Pearson
UA IHrekin	9,85	7,26	0,606
HTS IHrekin	9,03	6,74	0,741
Hibridoa IHrekin	8,19	6,11	0,752
Hibridoa IHrik gabe	9,89	6,99	0,699

2. taula. Pitch iragarpena hainbat metodorekin fonema ahostunetan erreferentzia naturalarekin konparatuz.

4.2. Ebaluaketa subjektiboak

Kutxa beltzeko esperimentu bat egin zen hiru modulu prosodiko konparatuz: UA (CART+RF iraupenerako eta UA pitcherako), HTS (iraupena eta pitcherako) eta IH. Guztietan modulu akustiko hibrido bera erabili zen. 27 pertsonak (6 aditu barne) lehentasuneko test batean parte hartu zuten ausazko albisteetako 10 esaldi balioztatuz. Naturaltasunari adituz, gehiago gustatzen zitzaien seinalea aukeratu behar zuten 5 balioko CMOS (Comparative MOS) eskalan: -2 (Lehenengo seinalean nabarmen nahiago dut) 2 (Bigarren seinalean nabarmen nahiago dut) bitartera. Emaitzak 2. irudian erakusten dira. IH metodo lehenetsia da 0,63 CMOS batekin versus HTS eta 0,31 versus UA. 95% KTk (Konfiantza Tartea) batez besteko tartea 0 gainetik jartzen du kasu bietan. Beraz, IH hobetsia da beste bi metodoekin alderatuta eta emaitzak estatistikoki adierazgarriak dira. Orokorrean, erantzunen %50,3k IH nahiago zuen, berriz, %19,7k baino ez zuen nahiago beste bi metodoetako bat.

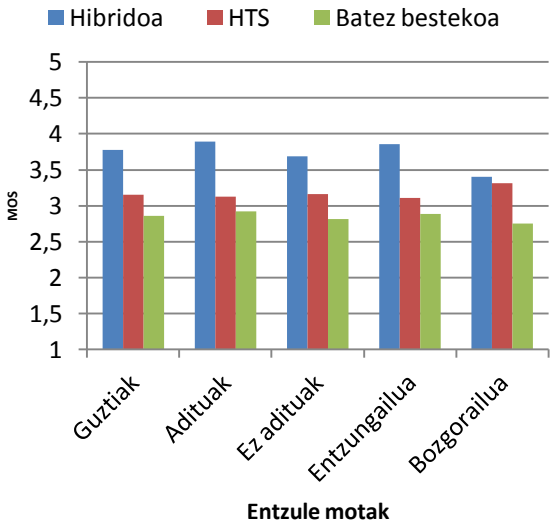
2. taulako iragarpen erroreak begiratuz harrigarria iruditu zitekeen UA prosodia eta IHren arteko ezberdintasun pertzeptuala HTSarekiko baino txikiagoa izatea. Lehen komentatu den bezala, UA modulu prosodikoa motore akustikoarekin batera diseinatua izan da eta argitu behar da entzuleak seinaleen naturaltasun orokorra balioztatzen ari zirela. Are gehiago, errealizazio prosodiko natural anitz daudenez eta entzuleek intonazioa guztira balioztatzen dutenez (eta ez lokalean) (Clark eta Dusterhoff , 1999), alderdi honi dagozkion emaitza objektiboak kontuz jaso behar dira.



2. Irudia: CMOS testa prosodiarako 3 metodo konparatuz, %95 KTarekin.

Albayzin 2010 TTS ebaluaketa kanpainara gure sistema hibridoa eta HTSan oinarritutakoa bidali genituen. Ebaluaketa honek gaztelaniazko datu-base berarekin garatutako sistema ezberdinak konparatzen ditu. Bere diseinua ingelesa eta txinatarrarentzako den Blizzard Challenge (Black eta Tokuda, 2005) kanpainan

oinarrituta dago. 132 entzulek parte hartu zuten ebaluaketa prozesuan. Gure sistema hibridoa ebaluaketaren hiru atazetan gailendu zen: jatorrizko ahotsarekiko antzekotasuna (4,07 MOS), naturaltasuna (3,71 MOS) eta ulergarritasuna (%17 WER). 3. irudiak naturaltasun emaitzak erakusten ditu bai sistema hibridoarentzat bai HTSen oinarritutakoarentzat, hainbat entzule-talderekin edo erabilitako ekipamenduarekin. Batez besteko sistema sintekiko guztien batez besteko puntuazioan datza. Wilcoxon testak erakutsi zuen gure sistema hibridoa besteak baino hobea zela naturaltasun-atazan adierazgarritasun estatistikoarekin. Bestalde, gainontzeko atazetan ez zegoen inolako ezberdintasun estatistikorik puntuazio onena zuten beste sistemekin. Emaitzei begiratuta ikusi daiteke entzungailuak erabili beharrean bozgorailuak erabiltzen badira, sistema hibridoa eta HTS arteko tartea nabarmen murrizten dela.



3. irudia: Naturaltasuna entzule ezberdinentzat.

5. Ondorioak

UA TTSk ahots naturala sintetizatzen dute eremu mugatuetan, baina zaratak eta distortsioak agertu ohi dira eremua zabaldu ahala. Naturaltasuna mantenduz trinkotasuna hobetzeko IH berri bat proposatu da. Bi iragarpen prosodiko konbinatzen dira eta HTS sistemaren aurresan espektrala UA modulu akustikoan erabiltzen da. Neurri objektiboez gain subjektiboek ere gure ikuspegiaren baliotasuna adierazten dute. IHk Aholaben UA TTSak batzuetan falta duen sendotasuna hobetzea lortu du. Bi iragarpen metodo konbinatzeak prosodia trinkoagoa sortu du (hau da, errore larri gutxiagorekin). Eta Aholaben modulu akustikoan EP TTSk sortutako parametro espektralak sartzeak unitate "txarren" aukeraketa murriztu du. Azalpen bera luzatu genezake ulergarritasun atazan izandako errendimendurako, non EP sistemak normalean emaitzarik onena lortzen duten datu-base txikiekin. *Albayzin 2010 TTS* ebaluaketaren emaitzei begiratuta, argi geratu behar du oraindik tarte nabarmena dagoela ahots naturala eta sintetikoen artean, zeren sistema sintetiko guztiek adierazgarriak ziren emaitza txarragoak lortu baitzituzten.

6. Esker onean

Lan honek Espaniako Zientzia eta Berrikuntza Ministerioaren (Buceador Proiektua, TEC2009-14094-C04-02) eta Eusko Jaurlaritzaren finantziaketa (Berbatek, IE09-262) jaso du.

7. Aipamenak

M. Aylett and C. Pidcock, "The CereProc Blizzard Entry 2009: Some dumb algorithms that don't work," in Blizzard Challenge Workshop, pp. 3-6, 2009.

A. Black, "Perfect synthesis for all of the people all of the time," in 2002 IEEE Workshop on, pp. 167-170, 2002.

A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in EUROSPEECH97, pp. 601-604, 1997.

A. Black and K. Tokuda, "The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets," in EUROSPEECH05, pp. 77-80, 2005.

A. Breen and V. Pollet, "Synthesis by Generation and Concatenation of Multi-Form Segments," in INTERSPEECH08, pp. 1825-1828, 2008.

L. Breiman, "Random forests," Machine learning, vol. 25, no. 2, pp. 5-32, 2001.

F. Campillo, J. van Santen, and E. Banga, "Integrating phrasing and intonation modelling using syntactic and morphosyntactic information," Speech Communication, vol. 51, no. 5, pp. 452-465, 2009.

R. Clark and K. Dusterhoff, "Objective methods for evaluating synthetic intonation," in EUROSPEECH99, pp. 1623-1626, 1999.

R. Donovan and P. Woodland, "Improvements in an HMM-based speech synthesiser," in EUROSPEECH95, pp. 573-576, 1995.

D. Erro et al., "HMM-based Speech Synthesis in Basque Language using HTS," in FALA2010, 2010.

D. Erro et al., "HNM-Based MFCC+F0 Extractor Applied to Statistical Speech Synthesis," in ICASSP11, 2011.

X. Gonzalvo et al., "High quality emotional HMM-based synthesis in Spanish," in ISCA Tutorial NOLISP, 2009.

I. Hernáez et al., "Description of the AhoTTS System for the Basque Language," in 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, 2001.

T. Hirai, J. Yamagishi, and S. Tenpaku, "Utilization of an HMM-based feature generation module in 5 ms segment concatenative speech synthesis," in IEEE Speech Synthesis Workshop, pp. 81-84, 2007.

HTSren Webgunea: "HMM-based Speech Synthesis System (HTS)." http://hts.sp.nitech.ac.jp/ (2012)

A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in ICASSP96, vol. 1, pp. 373-376, 1996.
H. Kawai et al., “XIMERA: A new TTS from ATR based on corpus-based technologies,” in 5th ISCA Workshop on Speech Synthesis, pp. 179-184, 2004.
S. Krstulovic, J. Latorre, and S. Buchholz, “Comparing QMT1 and HMMs for the synthesis of American English prosody,” in Speech Prosody, vol. 1, pp. 67-70, 2008.
A. Lazaridis et al, “Improving phone duration modelling using support vector regression fusion,” Speech Communication, vol. 53, no. 1, pp. 85-97, 2011.
I. Luengo et al., “Evaluation of Pitch Detection Algorithms Under Real Conditions,” in ICASSP07, pp. IV-1057-IV-1060, 2007.
F. Méndez et al., “The Albayzín 2010 Text-to-Speech Evaluation,” in Fala2010, pp. 317-340, 2010.
A. Raux and A. Black, “A unit selection approach to f0 modeling and its application to emphasis,” ASRU03, pp. 700-705, 2003.
S. Rouibia, O. Rosec, and T. Moudenc, “Unit Selection for Speech Synthesis Based on Acoustic Criteria,” in Text, Speech and Dialogue, pp. 281-287, 2005.
I. Sainz et al., “The AHOLAB Blizzard Challenge 2009 Entry,” in Blizzard Challenge 2009 workshop, 2009.
H. Silén et al, “Using Robust Viterbi Algorithm and HMM-Modeling in Unit Selection TTS to Replace Units of Poor Quality,” in INTERSPEECH10, pp. 166-169, 2010.
M. Siu and A. Chan, “A robust Viterbi algorithm against impulsive noise with application to speech recognition,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 6, pp. 2122-2133, 2006.
Y. Wu and R. Wang, “Minimum Generation Error Training for HMM-Based Speech Synthesis,” in ICASSP06, pp. I-89-I-92, 2006.
Yan, Z.-J., Qian, Y., and Soong, F.k., “Rich-context unit selection (RUS) approach to high quality TTS,” in ICASSP10, pp. 4798-4801, 2010.
J. Yang et al., “Multitier non-uniform unit selection for corpus-based speech synthesis,” in Blizzard Challenge Workshop, 2006.
H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” Speech Communication, vol. 51, no. 11, pp. 1039-1064, Nov. 2009.

