Ekaia, ale berezia 2021, 173-194
https://doi.org/10.1387/ekaia.22077

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN 0214-9001 – eISSN 2444-3255

Ahots sintetiko pertsonalizatuak: esperientzia baten deskribapena

(Personalised synthetic speech: description of an experience)

Inmaculada Hernáez Rioja, Eva Navas Cordón, Ibon Saratxaga Couceiro, Jon Sánchez de la Fuente*

HiTZ: Basque Center for Language Technologies - Aholab (UPV/EHU)

LABURPENA: Ahotsa ezinbestekoa da giza komunikaziorako, eta haren galerak eragin handia du pertsonak gizartean integratzeko prozesuan. Testu-ahots bihurketak ahots sintetikoa eman diezaieke ahozko desgaitasuna duten pertsonei. Irtenbide arruntenek ahots estandarra izaten dute normalean, eta, horregatik, erabiltzaile batzuek zailtasunak dituzte beren burua ahots horrekin identifikatzeko. Horregatik, ahots sintetiko pertsonalizatuak sortu behar dira, eta ahozko desgaitasuna duten pertsonei ahots-katalogo bat eskaini behar zaie, beren beharretara egokitzen den ahots bat aukeratu ahal izan dezaten. ZureTTS proiektuaren helburua ahots pertsonalizatu horiek ematea da, bai gaztelaniaz, bai euskaraz. Ahotsa galduko duten pertsonek edo ahotsik ez dutenei ahotsa eman nahi dieten pertsona altruistek 100 esaldi grabatzen dituzte, AhoMyTTS web-atariaren bidez. Esaldi horiekin, egokitze-prozesu bat egiten da, grabaketako ahotsaren antzeko ahots sintetiko bat sortzeko. Erabiltzaileari sintesi-motor bat ematen zaio ahots pertsonalizatu horrekin batera, ahozko mezuak sortzea eskaintzen duten aplikazioetan erabiltzeko. Gainera, ahots-katalogo bat ere badago, grabaketarik egin ezin duen pertsona batek ahots horien artean gustukoena aukeratu dezan. 1.200 pertsonak baino gehiagok erabili dute sistema hori ahots pertsonalizatu bat lortzeko, eta haietatik 58 hautatu ditugu katalogoan sartzeko. Erabiltzaileei egindako inkestek erakusten dute gustura daudela ahots sintetikoaren hainbat alderdirekin: gehienen ustez, ahots sintetikoa jatorrizkoaren antzekoa da, atsegina eta argia, baina robotiko samarra. Lan honek garapen jasangarrirako 10. helburuari laguntzen dio, herrialde bakoitzaren barneko eta herrialdeen arteko desberdintasunak murriztuz. Era berean, garapen jasangarrirako 4. helburuari ere laguntzen dio, guztiontzako kalitatezko hezkuntza inklusiboa nahiz bidezkoa bermatzea errazten duten tresnak eskainiz.

HITZ GAKOAK: ahots-sintesia; pertsonalizatutako ahots sintetikoak; gizaki-makina interfazeak; komunikazio alternatiboa eta handigarria.

ABSTRACT: *The voice is so essential for human communication that its loss drastically affects the integration of people in society. Text-to-speech can provide a synthetic voice for people with oral disabilities. The most common solutions usually provide a standard voice, and users have difficulties to identify themselves with it. For this reason, we need to create personalized synthetic voices and offer a catalogue of voices to people with oral disabilities so that they can choose one that suits their needs. The objective of the ZureTTS project is to provide these personalized voices, both in Spanish and in Basque. Through the AhoMyTTS web portal, people who are going to lose their voice or altruistic people who want to provide voices to those who do not have it, record 100 carefully selected sentences. A synthetic voice with similar characteristics to the voice of the recording is generated by applying an adaptation process. The user is provided with a synthesis engine along with that personalized voice, so that they can use it in applications that require oral message generation. In addition, we offer a catalogue of voices to choose from if one is no longer able to record. More than 1,200 people have used the system to obtain a personalized voice and 58 of them have been selected to be included in the catalogue. User surveys show user satisfaction with various aspects of the synthetic voice: most think that the synthetic voice is similar to the original, pleasant and clear, although a bit robotic. This work contributes mainly to goal 10 for sustainable development by reducing inequality within and among countries. It also contributes to goal 4 for sustainable development, providing tools that facilitate access for all to an inclusive, equitable and quality education.*

KEYWORDS: speech synthesis; personalized synthetic voices; human computer interfaces; Alternative and Augmentative Communication.

**** Harremanetan jartzeko / Corresponding author:*** Inmaculada Hernáez Rioja, HiTZ: Basque Center for Language Technologies - Aholab, University of the Basque Country (UPV/EHU), Bilboko Ingeniaritza Eskola I Eraikina, Torres Quevedo Ingeniaria Plaza, 1 (48013 Bilbo, Espainia UPV/EHU Bilbo. Euskal Herria). – inma.hernaez@ehu.eus – https://orcid.org/0000-0003-4447-7575.

Nola aipatu / How to cite: Hernáez Rioja, Inmaculada; Navas Cordón, Eva; Saratxaga Couceiro, Ibon; Sánchez de la Fuente, Jon (2021). «Ahots sintetiko pertsonalizatuak: esperientzia baten deskribapena»; *Ekaia,* ale berezia 2021, 173-194. (https://doi.org/10.1387/ekaia.22077).

Jasotze-data: 2020, urriak 05; Onartze-data: 2021, urriak 18

ISSN 0214-9001 - eISSN 2444-3255 /

SARRERA

Ahotsa gure komunikazio-tresnarik preziatuena da. Tamalez, gutako gehienok ez gara ohartzen haren balioaz, gure inguruko norbaitek (edo guk geuk) galdu arte. Orduan bakarrik konturatzen gara zein ezinbestekoa den hurbilekoekin komunikatzeko, etxean familiarekin edo lanean, tabernako barran edo terrazan. Egunean zehar hitz egiten ematen dugun denbora kontatuko bagenu, batez ere ahotsa lanerako tresna duten pertsonek, ziur asko harritu egingo ginateke emaitzekin.

Ahotsa edo hitz egiteko gaitasuna galtzea hainbat arrazoirengatik gerta daiteke: bat-batean (normalean, istripu bat izan eta fonazio-aparatua edo mintzamena sortzeaz arduratzen den nerbio-sistema kaltetzen direnean, edo lepoko minbiziaren osteko kirurgiaren ondorioz), edo pixkanaka. Horren adibide adierazgarriena AEA da (alboko esklerosi amiotrofikoa). Beste kasu batzuetan —garuneko paralisian, adibidez—, komunikatzeko zailtasunak lesio ez-progresiboen ondorio dira, eta arazo arinetatik ahozko ekoizpenaren ezintasunera artekoak izan daitezke. Garuneko paralisia duten haurrek komunikatzeko asmoa izan dezakete, baina hitz egiteko eta ulertarazteko ezintasunak haien komunikazio-trukea mugatzen du, eta horrek eragin negatiboa du haien progresio pertsonalean.

Garapen jasangarrirako 10. helburuak herrialdeen barneko eta herrialdeen arteko desberdintasunak murrizten ditu, eta inor atzean ez geratzea bermatzen du. Gure lana pertsona kalteberen kolektibo bati zuzenduta dago, eta gaixotze-arrazoiengatiko bazterkeria saihesteko tresna bat eskaintzen du. Gainera, tresna euskararentzat garatu da, eta horrek hizkuntza-arrazoiengatiko desberdintasuna murrizten laguntzen du.

Garapen jasangarrirako 4. helburuak (kalitatezko hezkuntza) honako helmuga hau du: «Guztiontzako kalitatezko hezkuntza inklusiboa nahiz bidezkoa bermatzea eta etengabeko ikaskuntzarako aukerak bultzatzea». Hezkuntza publikoko zerbitzuetarako sarbidea ez dago bermatuta adin guztietan komunikazio-zailtasunak dituzten pertsonentzat. 4.5 xedeak espliztuki hartzen ditu kontuan «desgaitasunen bat duten» pertsonak eta «ahultasun-egoeran dauden haurrak». Hala, 4.a atalean, proposatzen da «haurren eta desgaitasunen bat duten pertsonen beharrak eta genero-desberdintasunak kontuan hartzen dituzten eta ikaskuntza-ingurune seguruak, indarkeria gabeak, inklusiboak eta guztiontzako eraginkorrak eskaintzen dituzten hezkuntza-instalazioak eraikitzea eta egokitzea».

Teknologiak, hein batean, ahozko desgaitasuna murrizten lagundu dezake. Gaur egun, gai gara gizakiok ozen irakurtzeko dugun gaitasuna emulatzeko, TTS (Text-to-Speech) testu-ahots bihurketa izeneko teknologia erabiliz. Gaur eguneko ahots sintetikoak zailak dira giza ahotsetatik bereizten. Komunikazio Alternatibo eta Handigarriko aplikazioek (Augmentative and Alternative Communication, AAC) testu-ahots bihurketa beste tekno-

logia batzuekin konbinatzen dute (teklatua ordezten duten zentzumen-interfazeak edo mezuak azkar sortzeko teknikak), komunikazio-tresna eraginkorrak emateko.

Lan honetan, testu-ahots bihurketarako sistemei erreparatuko diegu, AAC teknologien funtsezko osagai gisa. Teknologiaren egungo egoera aurkezteaz gain, ahots sintetikoak eta, bereziki, ahots sintetiko pertsonalizatuak nola lortzen diren deskribatuko dugu. ZureTTS ekimena ere aurkeztuko dugu, euskarazko eta gaztelaniazko ahots sintetiko pertsonalizatuei sarbidea errazteko ekimena. ZureTTSn, ahots sintetiko propioa izateko aukera lehenetsi da beste neurketa-parametro batzuen aurrean, hala nola kalitatea edo naturaltasuna. Gure ahotsa gure nortasunaren parte da. Ahots sintetikoaren pertsonalizazioak komunikaziorako gailu elektronikoen erabilerak dakarren inpaktua murriztu nahi du.

Azkenik, ezin dugu ahaztu hizketa-teknologien garapen-maila desberdina dela munduko hizkuntzen artean. Hizkuntza minoritario edo gutxituetako hiztunentzat ere bermatu behar da AAC teknologietarako sarbidea.

TESTU-AHOTS BIHURKETA

Atal honetan, testu-ahots bihurketarako erabiltzen diren teknologien egoera deskribatuko dugu, ahots pertsonalizatuetarako bideko lehenengo urratsa baita. Lehenik, azken urteotan erabilitako eredu eta metodoak berrikusiko ditugu, haien arteko alde nagusiak azaltzeko. Ondoren, ahots-pertsonalizazioa lortzeko indarrean dauden estrategia batzuk azalduko ditugu. Ikerketa askok erakusten dute beste pertsona batzuen nortasunari buruzko iritzia osatzen dugula haien ahotsetik abiatuta (beste ezaugarri batzuekin gertatzen den bezala, hala nola aurpegiarekin edo azalaren kolorearekin) [1]. Ikerketa batzuek erakusten dute ahots pertsonalizatuak erabiltzeak garapen intelektuala erraztu diezaiekeela ikusmen-urritasunak dituzten haurrei [2]. Azken batean, ahots sintetikoen garapenak komunikazio-desgaitasuna duten pertsonak gizarteratzen laguntzen badu, ahots horien pertsonalizazioa lortzeak teknologia horien pertzepzioa hobetzen du, bai erabiltzailearen aldetik, bai inguruko pertsonen aldetik, eta, horrela, haien erabilera errazten da.

Azkenik, lan honi dagokion proiektuan inplementatutako sistemaren deskribapen bat egingo dugu atalaren amaieran.

Testu-ahots bihurketarako teknologiak

Testu-ahots bihurketaren helburua (Text to Speech Conversion, TTS) ahots naturalak sortzea da, hots, estilo jakin batean mintzatzeko eta giza hiztunen azentua, aldartea eta beste ezaugarri batzuk adierazteko gai direnak. Azken hamarkadaren hasiera arte, ahotsa sortzeko erabiltzen ziren teknologien

artean, unitate-hautaketaren bidezko sintesi kateatzailea [3] eta Markov-en ereduetan oinarritutako sintesi estatistiko-parametrikoa [4] erabiltzen ziren.

Unitate-hautaketaren bidezko sistema kateatzaileek aldez aurretik grabatutako ahots natural baten zatiak kateatuz sortzen dute ahotsa. Zatiak hautatzeko, irizpide konplexuak erabiltzen dira: alderdi akustikoak, fonetikoak, prosodikoak eta linguistikoak hartu behar dira kontuan. Orokorrean, unitate-hautaketan oinarritutako sistemek oso antzeko teknika bat aplikatzen dute ahotsaren intonazioa sintesi-prozesuan aurresateko [5]. Teknika horiekin, oso emaitza naturalak lortzen dira erabilera mugatutako giroetan, baina erabilera-domeinua zabaltzean lortzen diren emaitzen kalitatea oso aldagarria da [6]. Gainera, memoria-, biltegiratze- eta prozesatze-betebeharrak handiak dira, eta ahots berriak sortzeko malgutasuna, txikia.

Sistema estatistiko-parametrikoek, berriz, akustikoki antzekoak diren ahots-unitateen batez besteko ereduetan oinarrituz sortzen dute ahotsa. Vocoder baten bidez, hau da, ahotsa parametro akustiko bihurtzeko eta parametro horietatik seinalea birsortzeko gai den sistema baten bidez, ahotsa parametro sorta bat bihurtzen da [7]. Normalean, ahotsa zenbait parametro motatan deskonposatzen da: inguratzaile espektralarekin lotutako parametro espektralak, maiztasun-banda desberdinen energiari buruzko informazioa eramaten dutenak; intonazioarekin lotuta dagoen oinarrizko maiztasuna; eta iturriaren sonoritate-mailarekin erlazionatutako parametroak. Adibidez, banda desberdinen aperiodikotasunak STRAIGHT [8] eta WORLD [9] vocoderren kasuetan, edo gehieneko maiztasun ahostuna, AhoCoder vocoderraren kasuan [10]. Entrenamendu-datu kopuru bat gainditzen denean, sistema estatistiko-parametrikoek informazio fonetiko-linguistikoa eta dagokion vocoderrak ateratako parametroekin egindako errealizazio akustikoaren arteko erlazioa modelatzeko, Markov-en eredu ezkutuak erabiltzen dira (HMMs, Hidden Markov Models). Ondoren, sintesiaren momentuan, ereduek parametro akustikoen sekuentzia probableena itzultzen dute, sarreran dagoen testua deskribatzen duten etiketa fonetiko-linguistikoen sekuentziaren arabera. Sistema horien abantailen artean, trinkotasuna, malgutasuna, ulergarritasuna eta biltegiratze-betebehar txikia nabarmentzen dira. Ahots berriak erraz sortzeko aukera ematen dute, egokitze- edo interpolazio-tekniken bidez [11], eta ahots leuna eta kalitate egonkorrekoa sortzen dute, nahiz eta vocoderren erabilerak haien naturaltasuna murrizten duen. Haren ulergarritasuna hizkera naturalaren antzekoa da, eta are hobea giro zaratatsuetan [12].

Metodo bien abantailak konbinatzen dituzten hurbilketa hibridoak ere probatu dira. Batzuek estatistika-sistemaren aurresate prosodiko edota espektralak erabiltzen dituzte unitateak hautaketa-prozesuaren kostu objektiboa kalkulatzeko [13]. Beste batzuetan, sintesi kateatzailea erabiltzen da sistema estatistikoaren kalitatea hobetzeko [14], edo unitate naturalak eredu estatistikoek aurreikusitakoekin konbinatzen dira [6].

Azken urteotan, sintesi estatistiko parametrikoaren esparruan, Markov-en ereduen ordez neurona-sare sakonak (Deep Neural Networks, DNN) [15] erabiltzen ari dira, eta oso emaitza onak lortu dira ahots sintetikoaren kalitateari dagokionez. DNNak egokiak dira gaussiar ereduek ahotsa sortzeko parametro akustikoen eta hizketaren irudikapen sinbolikoaren arteko erlazio konplexu ez-linealen irudikapenean dituzten zenbait muga gainditzeko. Hainbat sare-arkitektura probatu dira, hala nola aurreranzko elikadura-sareak (feed-forward Networks)[16], sare errepikariak [17] eta WaveNet sareak[18]. Entrenamenduan erabilitako irizpideen artean, sortze-akats txikienarena nabarmentzen da bere aplikazioagatik [19], nahiz eta duela gutxi metodo berri bat proposatu den aurkako sortzaile-sareak erabiliz (GAN, Generative Adversarial Networks) [20], eta oso emaitza onak lortu dira ahotsaren naturaltasunari dagokionez. [21] lanean, ahotsaren parametro akustikoen sorreran sare sakonak erabiltzeko estrategia posibleen berrikuspen bikaina egin da.

Duela gutxi agertutako sistema batzuetan, neurona-sareek ez dute seinalearen sorreraren zatia soilik ordezkatzen: DNNen bidez, testua ahots bihurtzeko kate osoa egiten da. Deep Voice [22] TTS sisteman, etapa bakoitza neurona-sareen bidez inplementatu zen lehenengo aldiz. Lortutako seinalearen kalitatea ez da WaveNet bidez lortzen dena bezain ona, eta horregatik proposatu zen Deep Voice 2 [23] eta Deep Voice 3 [24] bertsioetan hobekuntzak egitea. Hala, seinalea WaveNet-en bidez sortzea lortu zen. Muturretik muturrerako (end-to-end) sistema baten antz handia duen hurbilketa bat Char2Wav [25] sistema da, nahiz eta oraindik parametroen iragarle-modulu bat eta vocoder neuronal bat eduki. Azkenik, TTSrentzat erabat muturretik muturrerakoak diren arkitekturak ere proposatu dira, hala nola Tacotron [26], Tacotron 2 [27] eta ClariNet [28], zeinek espektrogramak sortzen baitituzte testutik abiatuta. Ondoren, espektrograma horiek ahots bihurtzen dira, WaveNet edo Griffin-Limen algoritmoaren bidez [29]. TTS end-to-end ereduek bi osagai dituzte: kodegailu bat (enkoder) eta deskodegailu bat (decoder). Sarrerako sekuentziatik abiatuta (hitzak, karaktereak, fonemak eta baita byteak ere izan daitezke [30]), kodegailua irudikapen semantiko batean mapatzen saiatzen da, eta ezkutuko egoeren sekuentzia bat sortzen du. Deskodegailuak, egoera-sekuentzia hori arreta-mekanismo batekin testuinguru-informazio gisa erabiliz, deskodegailuaren ezkutuko egoerak eraikitzen ditu, eta irteera-tramak sortzen. Sistema horiek oso emaitza onak lortzen ari dira sortutako ahots sintetikoaren kalitateari dagokionez.

Ahots sintetiko pertsonalizatuak lortzeko teknikak

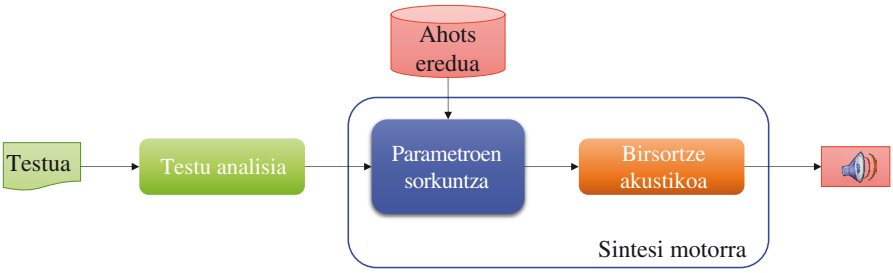
Orokorrean, TTS sistema estandarrek kalitatezko ahotsak eskaintzen dituzte, baina estandarrak: ezin da sistemaren ahotsa pertsonalizatuz ahots partikularrik garatu erabiltzaileen beharrak eta nahiak betetzeko. Neuro-

na-sareetan oinarritutako sistema sendo bat osatzeko behar den datu kopurua itzela da, eta, ahotsak egokitzeko eta sistemak datu kopuru murritzagoekin eraikitzeko esperimentuak egin diren arren [31], orokorrean ez da lortu kalitatezko ahotsak sortzea sistema estatistiko-parametrikoetan ahots pertsonalizatuak sortzeko erabil daitekeen bezain datu gutxirekin. Halako sistemetan, berriz, posible da egokitzapen-teknikak erabiltzea kalitatezko ahots sintetiko berriak sortzeko, datu kopuru mugatuarekin [32]. Horretarako, hasierako eredu estatistiko batzuk entrenatzen dira, eta, hala, hiztun anitzen datuekin, batez besteko ahotsa deritzona lortzen da. Batez besteko ahotsaren ereduetan, hizlariaren mendeko ezaugarriak edo beste ezaugarri espezifiko batzuk —generoa, adibidez— neutralizatu egiten dira, eta aldaera fonetikoa era sendoago batean modelatzen da. Horretarako, teknika espezifikoak erabiltzen dira; adibidez, hizlariari egokitutako entrenamenduaren bidezko parametroak berrestimatzea (SAT, Speaker Adaptive Training) [33]. Ondoren, desiratutako hiztun batek grabatutako 100etik 500era esaldi erabiliz, batez besteko ahotsaren ereduak moldatzen dira, bereziki diseinatutako teknikak erabiliz [34], hala nola gehieneko egiantzeko erregresio lineala (MLLR, Maximum-Likelihood Linear Regression) [35] edo gehieneko egiantzeko erregresio lineal mugatua (CMLLR, Constrained Maximum Likelihood Linear Regression) [36]. Lortutako ahotsak desiratutako hiztunaren ezaugarri bereizgarriak ditu, baina 100-500 esaldi horiek zuzenean ereduak entrenatzeko erabiliz lortuko litzatekeen baino kalitate eta sendotasun handiagoa. Hiztunari moldatzen zaizkion sistemak erabilgarriak dira, lanabes teknologiko gisa, ahotsean desgaitasun larriak dauzkatenei laguntzeko; laringotomia edo gaixotasun neurodegeneratiboak dituzten pazienteei, adibidez. Diagnostikoaren momentuan pazienteak egindako grabaketa batzuetatik abiatuta (haren ahotsaren narriadura oraindik gutxienekoa dela suposatuz), dagokion ahots artifiziala sor daiteke, sintetizadore baten bidez. Ildo horretako zenbait esperimentu nabarmen egin dira dagoeneko [37]. Teknologikoki, hiztunari egokitutako sistemek badute behar beste potentzial deskribatutako aplikazioari dagozkion atazak betetzeko, baina zenbait alor hobetzeko ikerketak egiten ari dira: ahotsaren parametrizazio-birsortzea [38], egokitzapen zuzena izateko material kantitatea [39], inguru zaratatsuetan grabatutako egokitzapen-datuen aurkako sendotasuna [40], hizkera sintetikoaren ulergarritasunaren hobekuntza ingurune bortitzetan [41], dagoeneko kaltetuta dauden ahotsen erabilera egokitzapenean erabiltzeko estrategiak [42], eta abar.

Ahots sintetikoaren pertsonalizazioan erabilitako teknika

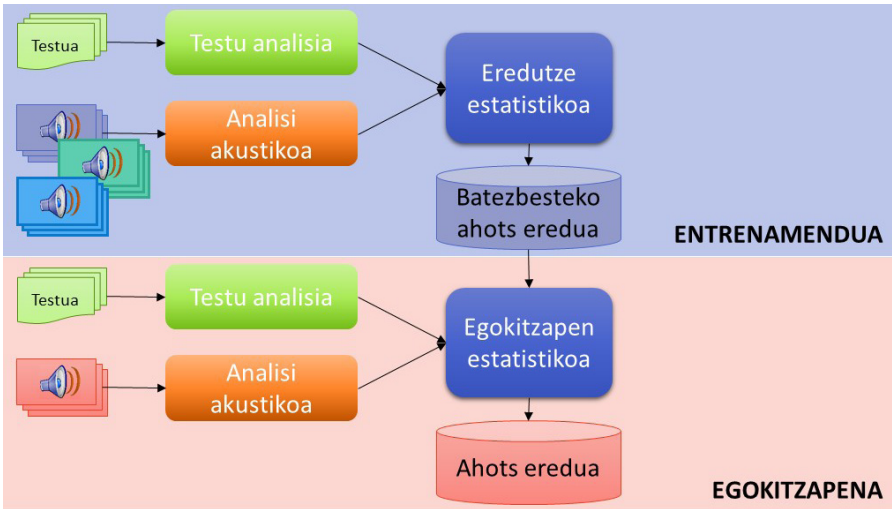
AhoLab laborategian garatutako testu-ahots bihurketarako sistemak, AhoTTS deritzonak [43, 44], TTS sistema estatistiko-parametrikoen egitura klasikoa erabiltzen du, **1. irudia**n ikus daitekeen moduan. Sarrera-testua prozesadore linguistiko baten bidez aztertzen da, sintesi-motorraren sarrera den informazio fonetiko-linguistikoa ateratzeko. Sintesi-motorrean, ahotsaren

eredu estatistikoak erabiltzen dira sarrerako etiketa fonetiko-linguistikoekin probabilitate handienarekin bat datozen vocoderraren parametroen balioak lortzeko. Azkenik, seinale sintetikoa vocoder baten bidez birsortzen da.



1. irudia. AhoTTS testu-ahots bihurketarako sistemaren egitura.

AhoTTSk testu-analisirako modulu bana dauka euskararako eta gaztelaniarako. Sintesirako erabiltzen diren ahots-ereduak ahots generiko batekin bat datoz, edo, AhoMyTTS atarian lortu bada, ahots pertsonalizatuarekin bat. Ahots pertsonalizatuaren ereduak sortzeko, **2. irudia**n erakusten den ahots-egokitzapenerako prozesua aplikatzen da:



2. irudia. Ahots egokituak sortzeko prozesua AhoMyTTSn.

Hiztun anitzeko grabaketetan eta ahoskatutakoari dagokion testuetan oinarrituta, eredu estatistiko bat garatzen da, AhoTTSren modulu linguis-

tikoaren bidez lortutako informazio fonetiko-linguistikoa vocoder batek grabaketetatik ateratako parametro akustikoekin erlazionatzeko. Eredu horrek batez besteko ahotsa irudikatzen du, hiztun desberdinei dagozkien aldaerak neutralizatzen baititu. Gaur egun, AhoMyTTSn erabiltzen den batez besteko ahotsa bi esatarik —gizonezkoa eta emakumezkoa— ahoskatutako 4.000 esaldirekin sortu da, bai gaztelaniaz, bai euskaraz.

Moldaketa-fasean, hiztun berriak kontu handiz aukeratutako 100 esaldi irakurtzen ditu, eta, moldatze-teknikak erabiliz, batez besteko ahotsaren ereduak aldatu, eta haren ahotsaren ezaugarriak erakusten dira. Ahots pertsonalizatuak sortu nahi diren hizkuntzako testu kopuru handietatik abiatuta hautatzen dira esaldien testuak; zehazki, haien transkripzio fonetikoa lortu, eta fonema isolatuen zein bi fonemen konbinazioen agerpen kopurua maximizatzen duten 100 esaldi hautatzen dira. Prozesu hori UPC unibertsitateak (Universitat Politècnica de Catalunya) sortutako CorpusCrt [45] lanabesa erabiliz gauzatzen da. Konbinazio fonetikoen kopurua maximizatzeko helburua dela eta, hautatutako esaldiek ezohiko soinuen konbinazioak izaten dituzte, eta, horren ondorioz, oso ohikoak ez diren hitzak dituen hiztegi bat sortzen da.

ZureTTS PROIEKTUA

Teknologiak ahots sintetiko pertsonalizatuak eman ditzakeela frogatu ondoren, hurrengo urratsa erabiltzaileengana iristea da. Ahots sintetiko pertsonalizatuak lortzeko, hainbat aukera komertzial daude, baina, guk dakigunaren arabera, guztiak ingeleserako [46, 47, 48, 49]. Oso posible da, baita ere, produktu eta ekimen berriak agertzea, hizkuntza-teknologiak indarra handia hartzen ari baitira merkatuan. Aipatutako ingelesezko sistema guztietan, «Voice Banking» terminoa erabiltzen da: hiztunaren ahotsa grabatzen da desgaitasuna nabarmena izan baino lehen, beharrezkoa den momentuan haren ahots sintetiko pertsonalizatua eskura eduki ahal izateko.

Halako sistemak euskaraz eskaintzeko aukera izan zen proiektu honetarako bultzada nagusia. 2011. urtean hasi zen, pertsonalizazioaren arloan ikertzen hasi ginenean [50]. Gaur egungo emaitza «AhoMyTTS» webataria da. Nahiz eta alderdi zailena garapen teknologikoa lortzea zela iruditu, egia esan, askoz zailagoa izan da erabiltzailearengana iristeko ezarpen praktikoa erdiestea. Helburu horrek funtsezko hiru alderdi eskatzen zituen:

1. Erabiltzailearen ahotsaren lagin multzoa lortzeko sistema sinple bat.
2. Erabiltzaileak ahots-sintesiaren motorra lortzeko metodo bat.
3. Emaileen ahotsen katalogo bat.

Lehenengoarekin, batez besteko ahotsa pertsonalizatzen da erabiltzaile emailearentzat, eta, bigarrenarekin, lortutako ahotsarekin komunikatzeko aukera ematen da. Hirugarren puntuarekin, hitz egin ezin duten pertsonei ahots pertsonalizatuak eskaintzea lortu dugu, hots, ahots pertsonalizatuak eskaintzen dizkiegu hitz egiteko gaitasuna galdu duten pertsonei. Horrek bereizi egiten gaitu aurretik aipatutako gainerako ekimenetatik.

Datozen ataletan, zehatzago azalduko dugu nola heldu diegun erronka horiei eta zein izan diren zailtasun nagusiak.

Emate-prozesua

Prozesu erraza da. Erabiltzaileari ahoskatu behar duen esaldia erakusten zaio; erabiltzaileak ahoskatzen du; behar bezala grabatuta dagoela egiaztatzen du, eta zerbitzarira bidaltzen du. Eskurapena esaldiz esaldi egiten da, prozesua zenbait saiotan osatu ahal izateko. Saio oso bat 30 eta 40 minutu artekoa izaten da.

Grabaketan, seinalearen maila egokia dela egiaztatzen du sistemak, batez ere saturazioak edo maila baxuegiak saihesteko.

Hauek dira ikusi ditugun arazorik garrantzitsuenak:

— Transmisioko erroreak konexio txar baten ondorioz.
— Ingurune zaratatsuetan egindako grabaketak (ate-kolpeak, haurren negarra, irratia edo telebista, edo trafiko-zarata).

Ildo horretan, interesgarria izan liteke hasieran grabaketa-ingurunea baliozkotuko lukeen metodoren bat sartzea.

Sintesi-motorra

Erabiltzaile batek grabazioak amaitu dituenean, ahotsa lortzeko prozesua aktiba daiteke. Prozesu hori erabat automatikoa eta opakua da erabiltzailearentzat. Haren eskaera aldizka kontsultatzen den itxarote-ilara batean sartzen da. Ahots bat egokitzeko prozesuak 40 minutu inguru behar ditu prozesadore estandar batean. Ondoren, erabiltzaileak mezu bat jasoko du, esteka batekin, ahotsa deskargatzeko (artikulu hau idazteko momentuan, sistemak Android eta Windowserako sintesi-motorra deskargatzeko aukera ematen du. Espero dugu iOS sistemetarako ere laster eskaini ahal izatea).

Ahots sintetikoen bankua

Ahots-katalogo batek ezaugarri desberdineko ahotsak erakutsi beharko lizkioke erabiltzaileari, ahots propiorik ez duen pertsonak bere burua hobe-

kien identifikatzen duena aukeratu ahal izan dezan (edo, besterik gabe, gustukoen duena). Erronka nagusia ahotsak nola erakutsi da, entzun behar direnak nolabait multzokatuta edo sailkatuta. Nahiz eta zenbait saiakera egin ditugun ezaugarri espektralak aztertuz eta inkestak eginez [51], emaitzak ez dira inoiz onak izan, eta, azkenean, metodo sinple bat aukeratu dugu, ahots bakoitzarekin edozein testu probatzeko aukera ematen duena, generoa aldez aurretik hautatuta.

Katalogoan eskaintzen diren ahotsak, noski, atariko emaileen ahotsak dira. Ahots guztiak ez dira sartzen; hortaz, aldez aurreko hautaketa bat egin da, eta, besteak beste, ulergarritasun-, argitasun- eta adierazkortasun-propietate onenak eskaintzen dituztenak aukeratu dira. Zoritxarrez, ez dago unibertsaltzat har daitekeen metodorik ahots sintetikoak ebaluatzeko. Testu-ahots bihurketarako sistemak ebaluatzeko, tradizionalki, alde batetik, ulergarritasuna neurtzen da (behar bezala identifikatutako hitzak edo silabak zenbatuz), eta, bestetik, adierazkortasuna edo naturaltasuna (normalean, lehentasunezko testa erabiliz). Horretarako, iritzia ematen duten giza entzuleak erabiltzen dira (gehien onartzen den neurria «batez besteko iritzi-puntuazioa» da, edo MOS, ingelesezko siglen arabera). Baina, praktikan, TTS baten ebaluazioak zer zereginetan erabiliko den hartu behar du kontuan: GPS baten gidaritza-ahots gisa [52], liburu bat irakurtzeko [53], edo komunikatzaile pertsonaletarako [54]. Horregatik guztiagatik, ulergarritasunaren eta kalitatearen ikuspegitik gutxieneko atalase bat gainditzen duten ahots guztiak sartzen ari gara ahots-bankuan, emaitzekin dugun esperientzian oinarritutako irizpide subjektibo bat erabiliz.

Artikulu hau idazteko unean, 7.016 pertsona zeuden sisteman erregistratuta (euskaraz eta gaztelaniaz), eta haietatik 1.245ek osatu dute grabaketa-prozesua eta sortu dute ahots sintetikoa (ikusi **1**. **taula**). Ez daukagu informazio zehatzik ahots hori zenbat jendek erabiltzen duen bere komunikazio-sisteman. Ahots horietatik guztietatik, 58 ahots-bankuan sartuta daude. Ahotsak aukeratzeko, kalitatearen irizpidea hartu da kontuan.

1. taula. AhoMyTTS sistemako erabiltzaile-, grabaketa- eta ahots kopuruak.

Hizkuntza	Erabiltzaile kopurua	Bukatutako grabaketak	Sortutako ahotsak	Ahots-bankuko ahots kopurua	
				Gizonezkoa	Emakumezkoa
Gaztelania	6.904	1.247	1.189	21	16
Euskara	112	63	56	11	10

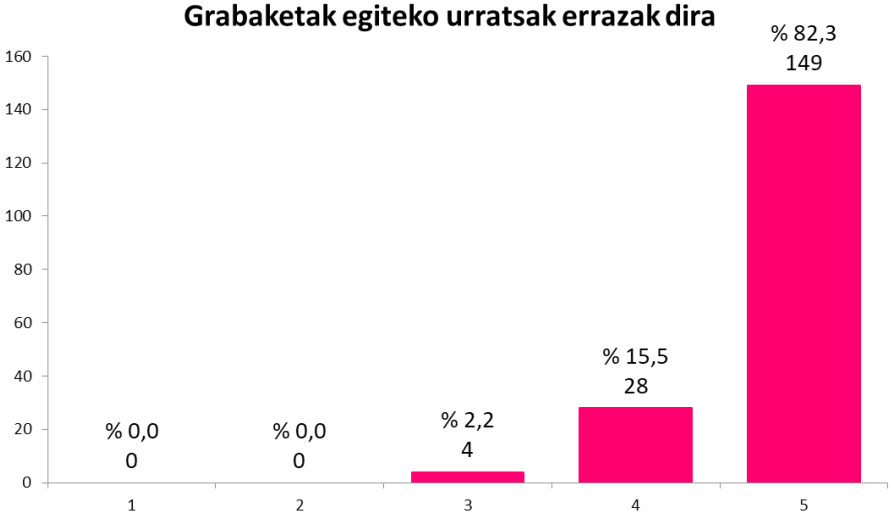
Emaitzak

Atal honetan, formulario baten bidez web-orrian bertan jasotako iritzien laburpena aurkeztuko dugu. Inkesta hainbat ataletan banatuta dago: webguneari buruzko galderak, emate-prozesuari buruzkoak eta lortutako azken ahotsari buruzkoak. Atal guztietan, erabiltzaileak 1etik 5erako eskala batean adierazi behar du erakutsitako adierazpenarekiko «adostasun» maila (1: Ez nago ados, 5: Erabat ados nago). Atal bakoitzean badago edozer gauza modu irekian adierazteko gune bat.

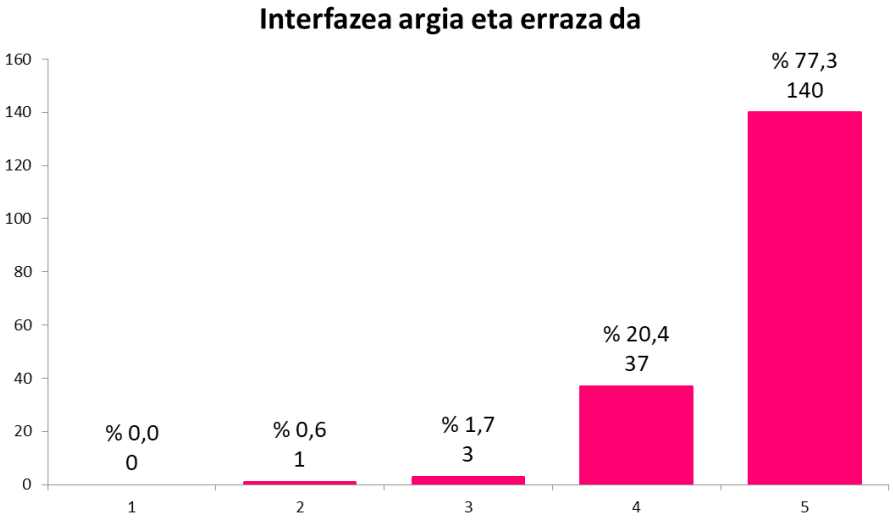
Garrantzitsua da inkesta erabat anonimoa dela azpimarratzea, eta erantzunak ez daudela lotuta sistemaren erabiltzailearekin. Hau da, ezin dugu jakin zer erantzun duen erabiltzaile jakin batek. Lan honetan, emate-prozesua egin duten 344 pertsonaren eta haien ahots sintetikoa lortu duten 198 pertsonaren erantzunak aurkeztuko ditugu.

Webgunea eta emate-prozesua

Inkesta bete duten erabiltzaile gehienen iritziz, interfazea argia eta erraza da, eta emate-prozesua erraza izan da, 3. eta 4. irudietan ikus daitekeenez.

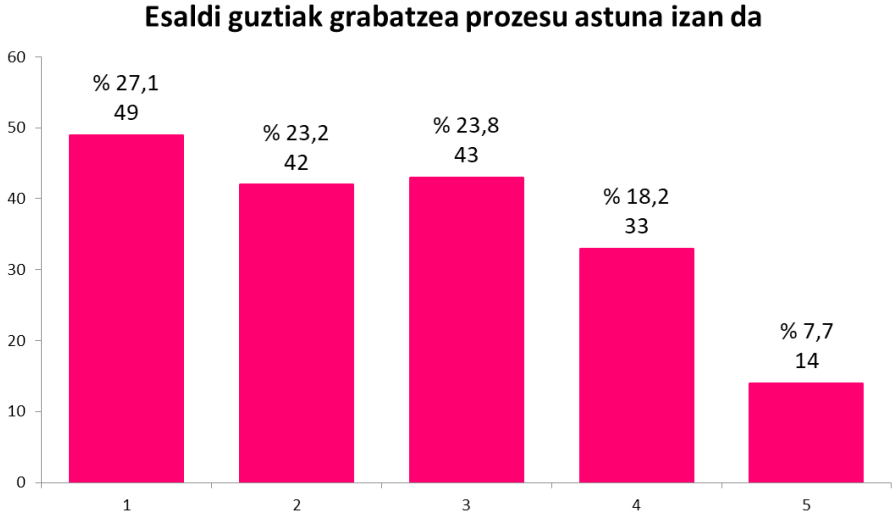


3. irudia. Erabiltzaileen iritzia grabaketa-prozesuari buruz.

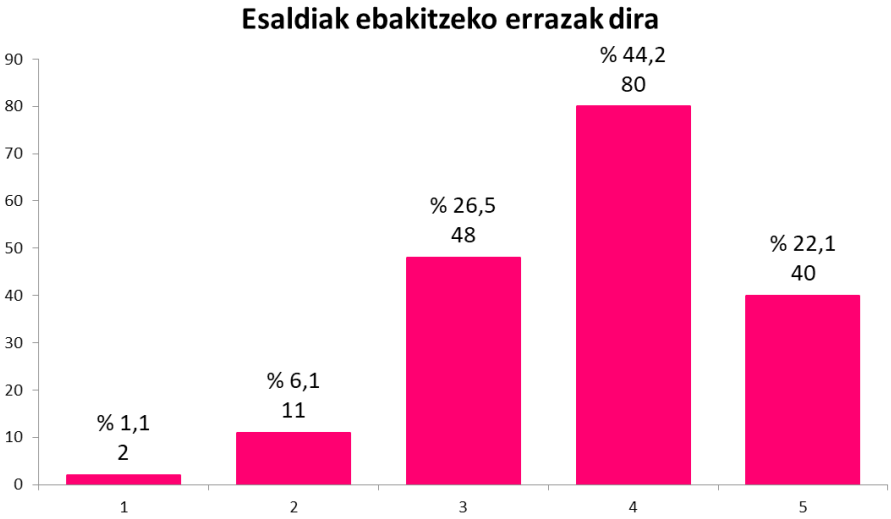


4. irudia. Erabiltzaileen iritzia interfazeari buruz.

Emate-prozesuari berari dagokionez, % 25,9k uste dute grabaketa-prozesua «astuna» izan dela. Hori oso bat dator pertsona gehienentzat esaldiak ahoskatzea erraza izatearekin (% 63,4). (5. eta 6. irudiak)



5. irudia. Erabiltzaileen iritzia prozesuaren astuntasunari buruz.

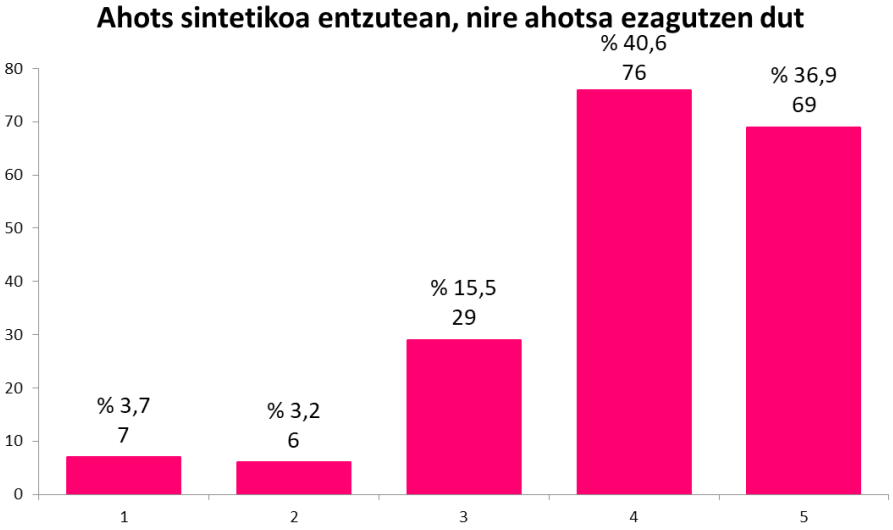


6. irudia. Erabiltzaileen iritzia esaldiak ahoskatzeari buruz.

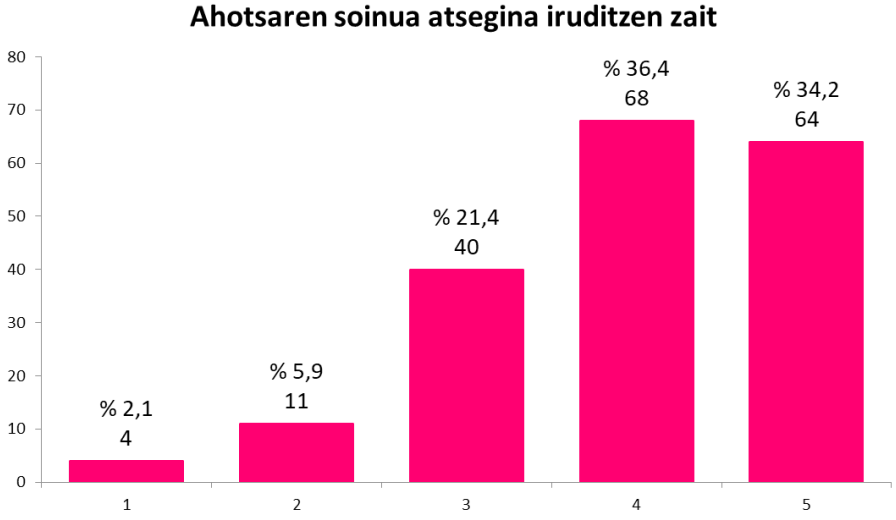
Ahots sintetikoak

Gure taldearentzat, emaitza pozgarrienetako bat da pertsonen % 76,4k beren ahotsa ahots sintetikoan ezagutzea (7. irudia). Gainera, oro har, inkestatutako pertsona gehienek uste dute moldatutako ahotsa «atsegina» (% 70,2) (8. irudia) eta adierazkorra (% 77,6) (9. irudia) dela, eta ahoskera argia dela (% 78,8) (10. irudia), baina, aldi berean, % 41,8k uste dute robotikoa ematen duela (11. irudia). Azken kasu horretan, halaber, pertsonen % 26,8k uste dute ahotsa ez dela robotikoa; beraz, inkestatuen heren batek (% 31,3) 3 batekin erantzun du (ez ados, ezta ez-ados ere).

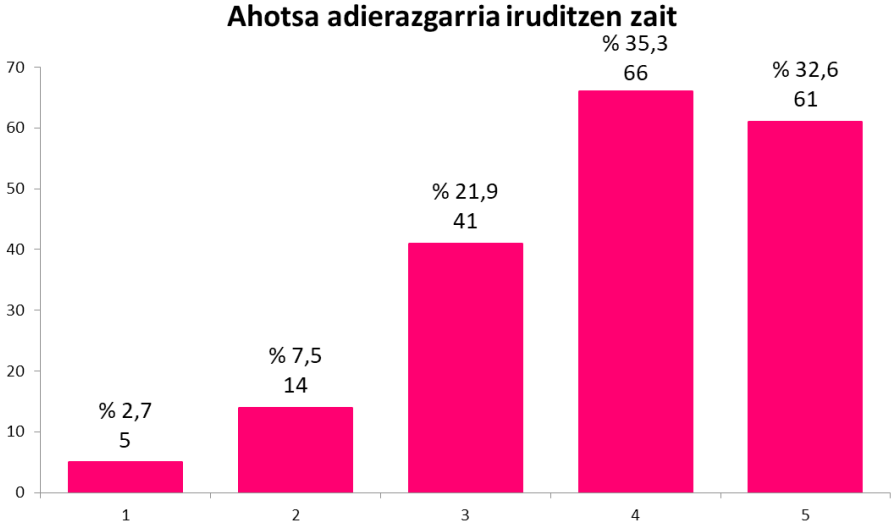
Azkenik, inkestak ahots sintetikoaren balorazio orokorrerako galdera bat jasotzen du, 1etik 10erako eskala erabiliz. 12. irudiko grafikoan ikus daitekeenez, oro har, asebetetze-maila oso handia da (% 73,2k 8ko edo gehiagoko balorazioa ematen diote, eta % 96k 5ekoa edo handiagoa).



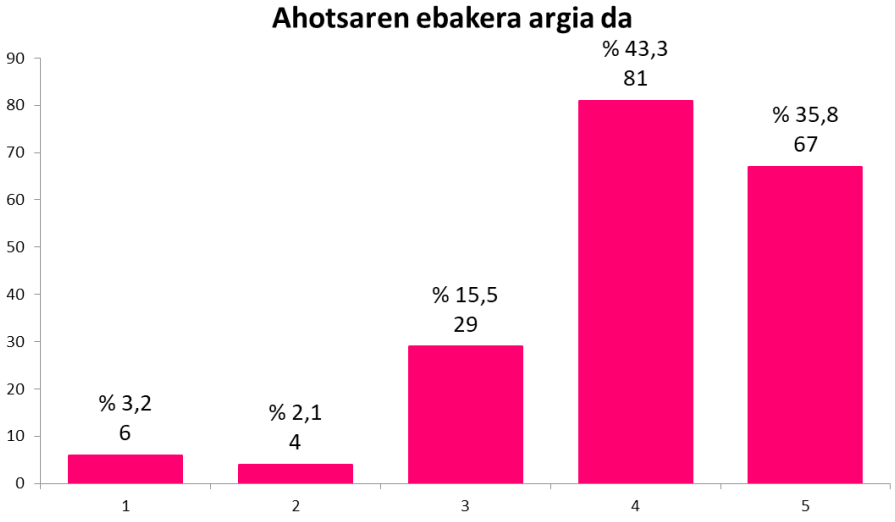
7. irudia. Erabiltzaileen iritzia ahotsa propiotzat hartzeari buruz.



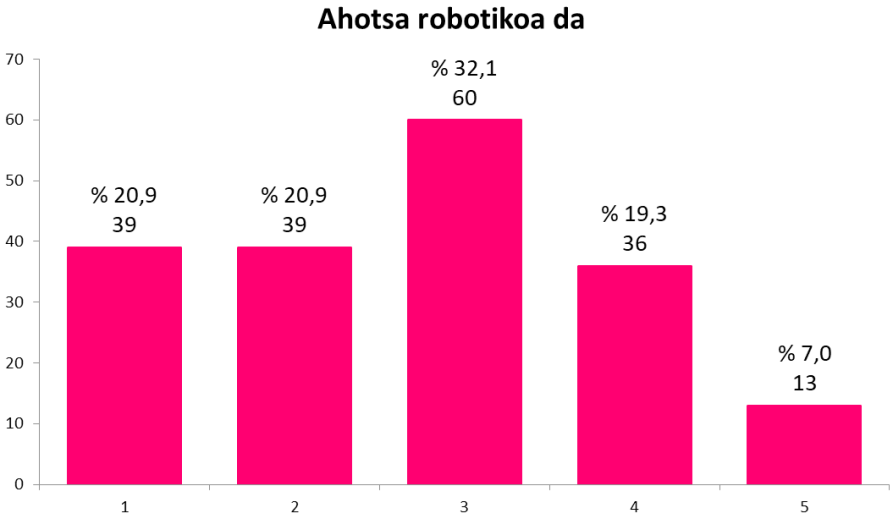
8. irudia. Erabiltzaileen iritzia ahotsaren atsegintasunari buruz.



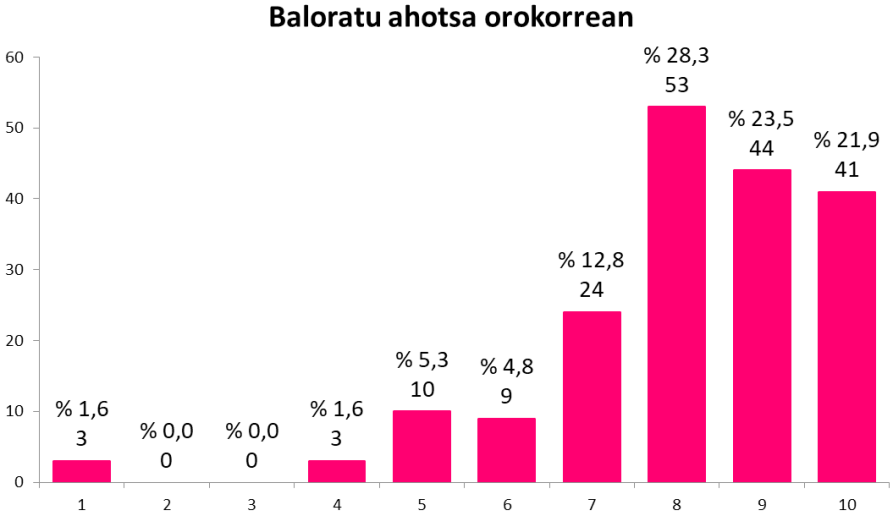
9. irudia. Erabiltzaileen iritzia adierazgarritasunari buruz.



10. irudia. Erabiltzaileen iritzia argitasunari buruz.



11. irudia. Erabiltzaileen iritzia robotikotasunari buruz.



12. irudia. Erabiltzaileen iritzi orokorra.

Erantzun irekiak

Inkestaren atal batean, erabiltzaileek proiektuaren webgunea hobetzeko eman dituzten iradokizunak jaso dira, era librean. Iradokizun gehienak (% 38 inguru) webgunearen erabilgarritasuna hobetzeari buruzkoak izan dira: nabigazioa sinplifikatzea eta erraztea; adibideak, jarraibideak edo erabiltzailearentzako feedbacka sartzea, eta horrelakoak. Iradokizun batzuek webgunean dagoeneko inplementatuta dauden funtzionalitateak eskatzen dituzte; beraz, nabigazioa ez da behar bezain intuitiboa, eta ez dute lortu horietara iristea.

Beste talde esanguratsu batek (% 12) hobekuntza estetikoak iradoki ditu orrialdeetan, bisualagoa eta diseinua biziagoa izan dadila eskatuz. Erabiltzaile kopuru berari (% 12) webgunea dagoen bezala ondo dagoela iruditzen zaio. Halaber, ikusmen-desgaitasuna duten pertsonentzat orrialdearen irisgarritasuna hobetzea eskatzen da (% 9), orrialdearen publizitatea handiagoa izatea eta ikusgarritasun handiagoa izatea (% 5), edo gailu mugikorretan erabili ahal izatea (% 3).

Gutxiagotan agertu dira webgunearekin, ahots pertsonalizatuekin eta ahots horiek lortzeko eta erabiltzeko prozesuarekin lotutako iradokizunak. Dena den, batzuek eskatu dute ematean esaldi sinpleagoak edo ohikoagoak erabiltzea (% 3), beste hizkuntza batzuk sartzea (% 3), katalogoko ahotsen ezaugarriak (dialektoa, azentua) etiketatzea (% 2) edo sortutako ahots sintetikoen kalitatea hobetzea (% 2). Iruzkinen batean, emandako ahotsen erabilerari buruzko xehetasun gehiago eskatu dira, eta emaile batek iradoki du emandako ahotsaren balizko hartzaileentzat mezu pertsonalak utzi ahal izatea. Informazio hori guztia oso kontuan hartuko da sistemaren etorkizuneko hobekuntzak diseinatzeko.

ONDORIOAK ETA ETORKIZUNERAKO LANAK

Lan honetan, ahots-sintesirako teknologien egoera deskribatu dugu, ahozko ezintasunak dituzten pertsonei laguntzeko euskarri gisa. Aisialdiaren eta entretenimenduaren munduan teknologia horiek merkatura azkar iritsi badira ere, desgaitasunerako aplikazioen munduan askoz motelagoa da aukera teknologikoen merkatura hurbiltzea. Garapen jasangarrirako 10. eta 4. helburuen ildotik, deskribatutako lanak hutsune hori bete nahi du, ahots sintetiko pertsonalizatua eskuragarriago egon dadin euskara- eta gaztelania-hiztunentzat.

Lanak arlo asko ditu hobetzeko oraindik. Garrantzitsuena neurona-sare sakonetan oinarritutako sintesi-teknologia erabiltzea da. Horrek nabarmen hobetuko luke ahotsen azken kalitatea, adierazkortasuna eta pertsonalizazioa bera barne.

Gure ustez garrantzitsua den beste alderdi bat aldaera dialektalak sartzea da, bai euskararentzat, bai gaztelaniarentzat, gaur egun aldaera «estandarra» besterik ez baita kontuan hartu. Euskararen kasuan, adibidez, Iparraldeko hiztunentzat egokiagoa den aldaera bat sar liteke. Guztiz egingarria da, lan honen egileek testu-ahots bihurketarako sistema bat garatu baitute dagoeneko aldaera horretarako [55]. Gaztelaniaren kasuan, badira hainbat baliabide akustiko hainbat eskualdetako azentua duten batez besteko ahotsak lortu ahal izateko, atarian erabiltzen den azentutik oso urrun daudenak, esate baterako, Andaluziako, Kanarietako edo Hego Amerikako aldaerak. Alderdi horrek nabarmen aberastuko lituzke ahots-bankuaren aukerak ere.

Interfazearekin lotuta hobetu beharreko alderdi teknikoen artean, itsuen irisgarritasuna hobetzea dago. Hobekuntza tekniko hori eta erabiltzaileek iradokitako beste batzuk, hala nola mugikorrean erabiltzeko aukera eta beste sistema eragile batzuetara hedatzea, pixkanaka garatuko dira.

Azkenik, azpimarratu nahi genuke lan hau jende askoren laguntzari esker izan dela posible: alde batetik, ikasle eta kolaboratzaile askok beren software-zatia garatu dute; bestetik, beren buruarentzat edo beste pertsona batzuentzat beren ahotsa eman duten pertsonen laguntza ordainezina da.

ESKER ONA

Lan hau Espainiako Ekonomia eta Lehiakortasun Ministerioren dirulaguntzaz (Spanish Ministry of Economy and Competitiveness with FEDER support, RESTORE project, TEC2015-67163-C2-1-R) eta Eusko Jaurlaritzaren dirulaguntzaz (Basque Government, DL4NLP KK-2019/00045, PIBA_2018_1_0035 eta IT355-19) egin da.

BIBLIOGRAFIA

[1] LAVAN, N., MILEVA, M., MCGETTIGAN, C. 2020. «How does familiarity with a voice affect trait judgements?». *British Journal of Psychology,* 112(1), 1-19.

[2] PUCHER, M., ZILLINGER, B., TOMAN, M., SCHABUS, D., VALENTINI-BOTINHAO, C., YAMAGISHI, J., SCHMID, E., WOLTRON, T. 2017. «Influence of speaker familiarity on blind and visually impaired children's and young adults' perception of synthetic voices». *Computer Speech and Language*, 46, 179-195.

[3] HUNT, A. J., BLACK, A. W. 1996. «Unit selection in a concatenative speech synthesis system using a large speech database». *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1, 373-376.

[4] ZEN, H., TOKUDA, K., BLACK, A. W. 2009. «Statistical parametric speech synthesis». *Speech Communication*, 51, 1039-1064.
[5] RAUX, A., BLACK, A. W. 2003. «A unit selection approach to F0 modeling and its application to emphasis». *2003 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2003,* 700-705.
[6] POLLET, V., BREEN, A. 2008. «Synthesis by generation and concatenation of multiform segments». *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,* 1825-1828.
[7] DUDLEY, H. 1939. «Remaking Speech». *Journal of the Acoustical Society of America*, 11, 169-177.
[8] KAWAHARA, H., MASUDA-KATSUSE, I., DE CHEVEIGNÉ, A. 1999. «Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds». *Speech Communication*, 27, 187-207.
[9] MORISE, M., YOKOMORI, F., OZAWA, K. 2016. «WORLD: A vocoder-based high-quality speech synthesis system for real-time applications». *IEICE Transactions on Information and Systems*, E99D, 1877-1884.
[10] ERRO, D., SAINZ, I., NAVAS, E., HERNÁEZ, I. 2014. «Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis». *IEEE Journal of Selected Topics in Signal Processing*, 8, 184-194.
[11] YAMAGISHI, J., USABAEV, B., KING, S., WATTS, O., DINES, J., OURA, K., TOKUDA, K., KARHILA, R., KURIMO, M. 2010. «Thousands of Voices for HMM-Based Speech Synthesis–Analysis and Application of TTS Systems Built on Various ASR Corpora». *IEEE Transactions on Audio, Speech and Language Processing*, 18, 984-1004.
[12] SUNI, A., RAITIO, T., VAINIO, M., ALKU, P. 2012. «The GlottHMM Entry for Blizzard Challenge 2012: Hybrid Approach». *Proc. of The Blizzard Challenge 2012*.
[13] SAINZ, I., ERRO, D., NAVAS, E., HERNÁEZ, I. 2011. «A Hybrid TTS Approach for Prosody and Acoustic Modules». *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,* 333-336.
[14] GONZALVO, X., GUTKIN, A., CARRIÉ, J. C., SANZ, I., TAYLOR, P. 2009. «Local minimum generation error criterion for hybrid HMM speech synthesis». *Proc. Interspeech,* 416-419.
[15] ZE, H., SENIOR, A., SCHUSTER, M. 2013. «Statistical parametric speech synthesis using deep neural networks». *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings,* 7962-7966.
[16] QIAN, Y., FAN, Y., HU, W., SOONG, F. K. 2014. «On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis». *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 3829-3833.
[17] FAN, Y., QIAN, Y., XIE, F.-L., SOONG, F. 2014. «TTS synthesis with bidirectional LSTM based Recurrent Neural Networks». *Proceedings of the Annual*

Conference of the International Speech Communication Association, INTERSPEECH, 1964-1968.

[18] OORD, A. VAN DEN, DIELEMAN, S., ZEN, H., SIMONYAN, K., VINYALS, O., GRAVES, A., KALCHBRENNER, N., SENIOR, A., KAVUKCUOGLU, K. 2016. «WaveNet: A Generative Model for Raw Audio». *arXiv preprint arXiv:1609:03499* (ikuste-data: 2021/10/01).

[19] WU, Z., KING, S. 2016. «Improving Trajectory Modelling for DNN-Based Speech Synthesis by Using Stacked Bottleneck Features and Minimum Generation Error Training». *IEEE/ACM Transactions on Audio Speech and Language Processing*, 24, 1255-1265.

[20] SAITO, Y., TAKAMICHI, S., SARUWATARI, H. 2018. «Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks». *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26, 84-96.

[21] LING, Z. H., KANG, S. Y., ZEN, H., SENIOR, A., SCHUSTER, M., QIAN, X. J., MENG, H., DENG, L. 2015. «Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends». *IEEE Signal Processing Magazine*, 32, 35-52.

[22] SERCAN O. ARIK, MIKE CHRZANOWSKI, ADAM COATES, GREGORY DIAMOS, ANDREW GIBIANSKY, YONGGUO KANG, XIAN LI, JOHN MILLER, ANDREW NG, JONATHAN RAIMAN, SHUBHO SENGUPTA, M. S. 2017. «Deep Voice: Real-time Neural Text-to-Speech». *International Conference on Machine Learning*, 195-204.

[23] ARIK, S., DIAMOS, G., GIBIANSKY, A., MILLER, J., PENG, K., PING, W., RAIMAN, J., ZHOU, Y. 2017. «Deep Voice 2: Multi-Speaker Neural Text-to-Speech». *Proc. Neural Information Processing Systems (NIPS)*, 2962-2970.

[24] PING, W., PENG, K., GIBIANSKY, A., ARIK, S., KANNAN, A., NARANG, S., RAIMAN, J., MILLER, J. 2017. «Deep Voice 3: 2000-Speaker Neural Text-to-Speech». *Proc. International Conference on Learning Representations (ICLR)*, 1-15.

[25] JOSE SOTELO, SOROUSH MEHRI, KUNDAN KUMAR, JOAO FELIPE SANTOS, KYLE KASTNER, AARON COURVILLE, Y. B. 2017. «Char2wav: End-to-end speech synthesis». *International Conference on Learning Representations*, 1-6.

[26] WANG, Y., SKERRY-RYAN, R. J., STANTON, D., WU, Y., WEISS, R., JAITLY, N., YANG, Z., XIAO, Y., CHEN, Z., BENGIO, S., LE, Q., AGIOMYRGIANNAKIS, Y., CLARK, R., SAUROUS, R. 2017. «Tacotron: Towards End-to-End Speech Synthesis». *Proc. Interspeech,* 4006-4010.

[27] SHEN, J., PANG, R., WEISS, R. J., SCHUSTER, M., JAITLY, N., YANG, Z., CHEN, Z., ZHANG, Y., WANG, Y., SKERRV-RYAN, R., SAUROUS, R. A., AGIOMVRGIANNAKIS, Y., WU, Y. 2018. «Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions». *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 4779-4783.

[28] PING, W., PENG, K., CHEN, J. 2018. «ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech». *arXiv preprint arXiv:1807.07281* (ikuste-data: 2021/10/01).

[29] GRIFFIN, D. W., LIM, J. S. 1984. «Signal Estimation from Modified Short-Time Fourier Transform». *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32, 236-243.

[30] LI, B., ZHANG, Y., SAINATH, T., WU, Y., CHAN, W. 2018. «Bytes are All You Need: End-to-End Multilingual Speech Recognition and Synthesis with Bytes». *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 5621-5625.

[31] TOMAN, M., MELTZNER, G. S., PATEL, R. 2018. «Data requirements, selection and augmentation for DNN-based speech synthesis from crowdsourced data». *Interspeech:Annual Conference of the International Speech Communication Association*, 2878-2882.

[32] YAMAGISHI, J., NOSE, T., ZEN, H., LING, Z.-H., TODA, T., TOKUDA, K., KING, S., RENALS, S. 2009. «Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis». *IEEE Transactions on Audio, Speech, and Language Processing*, 17, 1208-1230.

[33] ANASTASAKOS, T., MCDONOUGH, J., MAKHOUL, J. 1997. «Speaker adaptive training: A maximum likelihood approach to speaker normalization». *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1043-1046.

[34] YAMAGISHI, J., KOBAYASHI, T., NAKANO, Y., OGATA, K., ISOGAI, J. 2009. «Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm». *IEEE Transactions on Audio, Speech, and Language Processing*, 17, 66-83.

[35] TAMURA, M., MASUKO, T., TOKUDA, K., KOBAYASHI, T. 2001. «Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR». *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 805-808.

[36] DIGALAKIS, V. V., RTISCHEV, D., NEUMEYER, L. G. 1995. «Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures». *IEEE Transactions on Speech and Audio Processing*, 3, 357-366.

[37] CREER, S. Personalising Synthetic Voices for Individuals with Severe Speech Impairment, Ph.D. Disertation, University of Sheffield, 2009.

[38] ERRO, D., SAINZ, I., NAVAS, E., HERNÁEZ, I. 2011. «Improved HNM-Based Vocoder for Statistical Synthesizers». *Interspeech*, 1809-1812.

[39] ERRO, D., ALONSO, A., SERRANO, L., NAVAS, E., HERNAEZ, I. 2013. «New Method for Rapid Vocal Tract Length Adaptation in HMM-based Speech Synthesis». *Eighth ISCA Workshop on Speech Synthesis*, 125-128.

[40] YANAGISAWA, K., LATORRE, J., WAN, V., GALES, M. J. F., KING, S. 2013. «Noise Robustness in HMM-TTS Speaker Adaptation». *Proc. 8th ISCA Speech Synthesis Workshop*, 119-124.

[41] ERRO, D., ZORILĂ, T. C., STYLIANOU, Y., NAVAS, E., HERNÁEZ, I. 2013. «Statistical synthesizer with embedded prosodic and spectral modifications to generate highly intelligible speech in noise». *Proc. Interspeech*, 3557-3561.

[42] YAMAGISHI, J., VEAUX, C., KING, S., RENALS, S. 2012. «Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction». *Acoustical Science and Technology*, 33, 1-5.

[43] SAINZ, I., ERRO, D., NAVAS, E., HERNÁEZ, I., SÁNCHEZ, J., SARATXAGA, I., ODRIOZOLA, I., LUENGO, I. 2010. «Aholab Speech Synthesizers for Albayzin2010». *VI Jornadas de Tecnologías del Habla and II Iberian SL Tech Workshop FALA 2010*, 343-348.

[44] HERNAEZ, I., NAVAS, E., MURUGARREN, J. L., ETXEBARRIA, B. 2001. «Description of the AhoTTS Conversion System for the Basque Language». *SSW4-2001*, 202.

[45] SESMA, A., MORENO, A. 2000. *CorpusCrt 1.0: Diseño de Corpus Orales Equilibrados*. Technical Report, UPC.

[46] Model Talker, https://www.modeltalker.org/ (ikuste-data: 2021/10/01).

[47] The Voice Keeper, https://thevoicekeeper.com/ (ikuste-data: 2021/10/01).

[48] VocalID, https://vocalid.ai/ (ikuste-data: 2021/10/01).

[49] SpeakUnique https://www.speakunique.co.uk (ikuste-data: 2021/10/01)./

[50] ERRO, D., HERNÁEZ, I., ALONSO, A., GARCÍA-LORENZO, D., NAVAS, E., YE, J., ARZELUS, H., JAUK, I., HY, N., MAGARIÑOS, C., SULÍR, M., TIAN, X., WANG, X., PEREZ RAMON, R. 2015. «Personalized Synthetic Voices for Speaking Impaired: Website and App». *Proc. Interspeech*, 2015, 1251-1254.

[51] ARRUTI, J. Ahots Sintetikoak Aukeratzeko Katalogoaren Diseinua, master amaierako lana, Euskal Herriko Unibertsitatea, 2017.

[52] LO, E.-W. (VICTOR), GREEN, P. 2013. «Development and Evaluation of Automotive Speech Interfaces: Useful Information from the Human Factors and the Related Literature». *International Journal of Vehicular Technology*, 2013.

[53] HINTERLEITNER, F., NEITZEL, G., MÖLLER, S., NORRENBROCK, C. R. 2011. «An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks». *Blizzard Challenge Workshop*.

[54] FIANNACA, A. J., PARADISO, A., CAMPBELL, J., MORRIS, M. R. 2018. «Voicesetting: Voice authoring uis for improved expressivity in augmentative communication». *Conference on Human Factors in Computing Systems - Proceedings*, 1-12.

[55] NAVAS, E., HERNAEZ, I., ERRO, D., SALABERRÍA, J., OYHARÇABAL, B., PADILLA, M. 2015. «Nafar-Lapurtar euskalkiarentzako euskal TTS bat garatzea». *Euskalingua*, 26, 22-27.