EKAIA (2024), artikulua prentsan/article in press
https://doi.org/10.1387/ekaia.26334
Behin-behineko bertsioa (euskara-orrazketaren faltan).

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN: 0214-9001
e-ISSN: 2444-3255
©UPV/EHU Press

Hizlari-bektore manipulazioaren bidezko genero-anbiguoko hizketaren sintesia euskaraz

(Gender-ambiguous speech synthesis in Basque using speaker embedding manipulation)

*Xabier Sarasola**, *Ander Corral, Igor Leturia, Iñigo Morcillo*
Orai NLP Teknologiak, Usurbil

LABURPENA: Gero eta interes handiagoa dago genero-anbiguko ahotsa duten text-to-speech (TTS) sistemetan, besteak beste, laguntzaile birtualetan eta bozgorailu adimendunetan genero-alborapenak eta estereotipoak saihesteko duten ahalmenagatik. Artikulu honetan, ahots-bihurketa teknika berriak aplikatzen dizkiegu ahots-bektoreei sare neuronaletan oinarritutako genero-anbiguoko euskarazko TTS sistemak lortzeko. Hizlari-bektoreak hiztun anitzeko Tacotron 2-a entrenatuz lortzen dira. Hizlari-bektoreen normalizazioa eta eskala parametro bat erabiltzen duten eta erabiltzen ez duten sistemak konparatu ditugu, baita genero bakoitzaren batez besteko bektore eta ahots errealen bektoreetan erabilera sistema horietan. Emaitzek frogatzen dute aurkeztutako metodoak genero-anbiguoko ahotsak lortzeko baliozkoak direla kalitate onargarriak, baina hobetu daitezkenak, lortuz.
HITZ GAKOAK: hizketaren sintesia, genero-anbiguoko ahotsa, hizlari-bektoreak, ahots transformazioa, euskara.

ABSTRACT: *There is a growing interest in text-to-speech (TTS) systems with gender-ambiguous voices, among other things due to their potential to avoid gender biases and stereotypes in voice assistants and smart speakers. In this paper we present and evaluate some novel methods that apply voice morphing techniques to speaker embeddings in order to obtain neural network-based gender-ambiguous voiced TTS systems for the Basque language. The speaker embeddings are obtained training a multi-speaker Tacotron 2. We compare the performance of systems with and without speaker embedding normalization with a scaling parameter, and also the application of these systems to the average embeddings of each gender and to real voice embeddings. The results prove that the methods presented are valid to obtain gender-ambiguous voices with acceptable, albeit improvable, quality.*
KEYWORDS: speech synthesis, gender-ambiguous voice, speaker embeddings, voice morphing, Basque.

***Harremanetan jartzeko/Corresponding author:** Xabier Sarasola, Orai NLP Technologies, Basque Country.
https://orcid.org/0000-0002-9220-5075, x.sarasola@orai.eus

Nola aipatu/How to cite: Xabier Sarasola; Ander Corral; Igor Leturia; Iñigo Morcillo, 2024, «Hizlari-bektore manipulazioaren bidezko genero-anbiguoko hizketa sintesia euskaraz», Ekaia, DOI: https://doi.org/10.1387/ekaia.26334

Jasoa: maiatzak 17, 2024; Onartua: uztailak 22, 2024
ISSN 0214-9001-eISSN 2444-3225 / ©2024 UPV/EHU

1. Sarrera

UNESCOk txosten bat argitaratu zuen 2019an [1], non laguntzaile birtualetan dagoen sexismoa nabarmendu zen eta laguntzaile birtual feminizatuak erabiltzearen ondorio kaltegarri ugari aipatzen ziren. Txostenean 18 gomendio ematen dira generoari buruzko joera horiek prebenitzeko. Horietatik 7.ak proposatzen du ez gizonezkoa ez emakumezkoa den laguntzaile birtual baten bideragarritasuna aztertzea. Gainera, beranduago frogatu da ahots horiek areagotu egiten dutela emakumeen laguntzaile birtualekiko konfiantza, haiekiko konfiantza oro har kaltetu gabe [2].

Genero-anbiguoko ahotsa lortzeari buruzko literatura urria da, eta lan gehienek oinarrizko frekuentziak entzuleek generoan hautematen dituzten efektuak aztertzen dituzte [2], adibidez Q lehen genero-anbiguoko ahotsean bezala [1] [3]. Lan horietan, ahots natural eta oinarrizko frekuentzia aldatuko ahots asko entzule ugarirekin ebaluatzen dira, oinarrizko frekuentzia edo ahots-ezaugarri anbiguoenak zein diren ikasteko. Lan honetan, hizlari-bektoreen manipulazioa proposatzen dugu generoaren anbiguotasuna lortzeko. Ahotsaren-bektoreen manipulazioaren bidezko ahots bihurketak honako premisa hau hartzen du kontuan: hizlari anitzeko sintesi sistema bidez lortutako hizlari-bektoreek hizlarien ezaugarriak adierazten dituen espazio latente bat ordezkatzen dutela. Hizlari-bektoreak erabiltzen dituzten hizlari anitzeko sistemak oso erabiliak dira gaur egun hainbat hizlariren ahotsa sintetiza dezaketelako eredu bakarraren bidez eta hizlari bakoitzeko grabazio gutxiago behar dituztelako honetarako [4, 5]. Hala ere, ahots-bektoreen manipulazio bidez ahots berriak sortzea ez da izan asko landu den ikerketa-lerroa. Gure inspirazioa Arik et al.-en lanetik dator [6]; hitz-bektoreetan aplikatzen diren manipulazioak [7] aplikatzen dituena hizlari-bektoreetan genero eta azentu transformazioak (ingeles britaniarretik estatubatuarrera eta alderantziz) lortzeko. Gure kasuan, erdibideko genero-eraldaketak egin ditugu, genero-anbiguotasuna lortzeko. Gainera, metodo berri bat proposatzen dugu hizlari-bektoreak entrenatu eta manipulatzeko, distantzia angeluarreko metodoek hizlarien ezagutzaren artearen egoeran izan duten arrakasta kontuan hartuta [8]. Sortu ditugun ahots sintetikoen adibideak demo webgunean entzun daitezke[2].

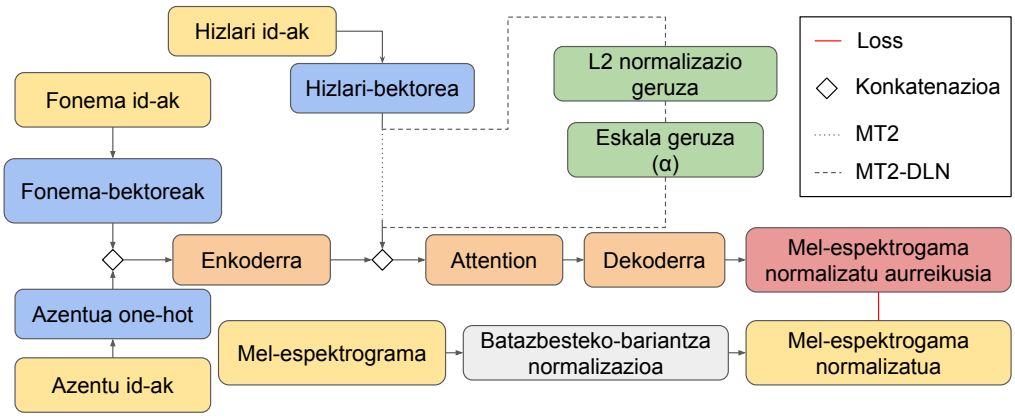
2. Hizlari aniztun hizketaren sintesi eredua

Gure hizketa-sintesi sistema ohikoak diren bi etapatan banatuta dago: parametroen sorkuntza eta vocoderra. Parametro-sorkuntzak hizlari-bektore taula erabiltzen duen Tacotron 2 hizlari-anitzaren egitura dauka [9] baina aldaketa batzuk gehitu dizkiogu. Honez gain, genero-anbiguoko ahotsak sortzeko bi teknika ezberdin probatu ditugu, MT2 eta MT2-DLN deitu diegunak, eta datozen azpiataletan deskribatuko ditugunak. Gure ereduak NVIDIAren Tacotron 2 biltegian oinarrituta daude [3]. Waveglow erabili dugu vocoder bezala [10] NVIDIAk partekatutako inplementazioa erabiliz[4].

2.1. MT2

MT2 teknika hizlari-bektoreek baldintzatutako Tacotron 2 hizlari-anitzean oinarritzen da [9], baina sarrera eta irteera informazioa eta sare neuronalaren galera aldatu ditugu. Honez gain fonema sekuentziak erabiltzen ditugu sarreran eta fonema bakoitzean azentuaren presentzia konkatenatzen da 2-Dko one-hot bektore baten bidez [11]. Irteeran, zero batezbesteko eta unitate bariantza normalizazioa aplikatzen diogu mel-espektrogramari entrenamenduko mel-espektrograma erreferentziatzat hartuz. Inferentzian, sarearen irteera denormalizatzen da vocoderrari pasa aurretik. L_2 *loss*-a bakarrik erabiltzen dugu jatorrizko hizlari-bakarreko Tacotron 2-an bezala. Sistemaren eskema orokorra 1. Irudian ikus dezakegu.

[1] https://genderlessvoice.com
[2] https://orai-nlp.github.io/Genderless-demos
[3] https://github.com/NVIDIA/tacotron2
[4] https://github.com/NVIDIA/waveglow



1. irudia. MT2 eta MT2-DLN-ren egitura eskema.

2.2. MT2-DLN

MT2-DLN teknika MT2 sistema bezalakoa da baina Deep Length Normalization (DLN) aplikatzen dugu L2 normalizazio geruza eta ikasgarria den eskala parametro baten bidez [12]. Honen helburua hizlari-bektore guztiak tamaina ikasgarria duen hiperesfera batean kokatzean datza. Sarean erabilitako bektoreen luzera berdina denez, luzera normalizatuaren bertsioan egin diezazkiegu transformazioak ondoren berriz sarean erabiltzeko. Honek esklusiboki distantzia angeluarrekin lan egitea ahalbidetzen digu. Sarearen eskema 1. Irudian ikus daiteke, pauso gehigarriak markaturik daudela.

3. Hizlari-bektoreen manipulazioa

Hainbat operazio aplikatzen zaizkie entrenatutako bektoreei genero-anbiguoko bektoreak kalkulatzeko. Bi mekanismo desberdin probatu ditugu:

- Gizonezkoen eta emakumezkoen batezbesteko bektoreak kalkulatu eta bien artean dagoen bektorea bilatu.
- Ahots erreal baten bektorea manipulatu genero bakoitzaren batezbesteko bektoreen artean dagoen genero-anbiguetate espaziora mugitzeko.

Proposatzen dugun sistema bakoitzean operazio hauek ezberdinak dira, MT2an distantzia euklidearrak eta MT2-DLNan distantzia angeluarrak kontsideratzen ditugulako. Operazio bakoitza ondoko azpiataletan azalduko dugu.

3.1. MT2 bektore manipulazioa

MT2 sisteman, [6]-ko logika euklidearra jarraitzen dugu bektore berriak lortzeko. Genero bakoitzaren batez besteko bektorea kalkulatzen dugu genero bakoitzaren zentroide bezala, ondoko formularekin:

$$\boldsymbol{c} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \tag{1}$$

non $\boldsymbol{x}_i$ aukeratutako generoko edozein hizlariren bektorea den eta N genero horretan dauden hizlari kopurua den. Ondoren, genero-anbiguoko zentroidea ($\boldsymbol{c}_{GA}$) kalkulatzen dugu ondoko moduan:

$$\boldsymbol{c}_{GA} = \boldsymbol{c}_G + \frac{(\boldsymbol{c}_E - \boldsymbol{c}_G)}{2} \tag{2}$$

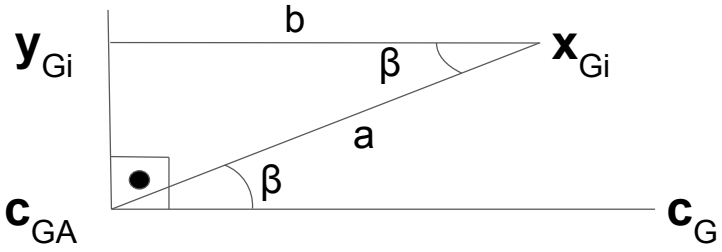
non $\boldsymbol{c}_G$ gizon zentroidea den eta $\boldsymbol{c}_E$ emakume zentroidea den.

Edozein hizlari-bektore genero-anbiguoko bektore batean bihurtzeko transformazio bektorearen norabidea eta modulua kalkulatu behar dugu. Transformazio bektorearen norabidea $\boldsymbol{c}_E - \boldsymbol{c}_G$ da [6], baina gure helburua transformatutako bektorea gizon eta emakume zentroideetara distantzia euklidear bera duen posizio batera mugitzea da. Puntu hau genero-anbiguoko zentroidea duen eta genero zentroideen arteko norabidearen ortogonala den planoan kokatuta dago. Hizlari-bektorea plano honetara eramaten duen modulua trigonometria aplikatuz (ikus 2. irudia) lor dezakegu, ondoko operazioaren bidez:

$$b = a \cos \angle(\boldsymbol{x}_{Mi} - \boldsymbol{c}_{GA}, \boldsymbol{c}_M - \boldsymbol{c}_{GA}) \tag{3}$$

$$\boldsymbol{y}_{Mi} = \boldsymbol{x}_{Mi} + b\frac{\boldsymbol{c}_E - \boldsymbol{c}_G}{\|\boldsymbol{c}_E - \boldsymbol{c}_G\|} \tag{4}$$

non $\boldsymbol{y}_{Gi}$ hizlari-bektore bihurtua den eta $\boldsymbol{x}_{Gi}$ gizon-bektore originala den. Gainontzeko aldagaiak 2. Irudian definituak daude. Emakume hizlari-bektorearen bihurketan ($\boldsymbol{x}_{Ei}$), genero marka guztiak alderantzizkatzen dira formulan. Inferentzia garaian, genero-anbiguoko bektorea eta transformatutako bektoreak erabiltzen dira genero-anbiguoko hizlarien mel-espektrogramen sintesia egiteko.



2. irudia. Transformazioa euklidear espazioan.

3.2. MT2-DLN bektore manipulazioa

MT2-DLN sisteman, bektoreak normalizatzen dira sarean sartu haurretik. Honek esan nahi du entrenatutako bektore guztiak normalizatu ditzakegula eta unitate luzerako hiperesfera batean egin ditzakegula transformazio guztiak. Honek transformazioak distantzia angeluarretan egiten baimentzen digu. Luzera normalizatua daukaten hainbat bektoreren batez besteko angelua ondoko eran lor daiteke:

$$\boldsymbol{c\prime} = \frac{1}{N}\sum_{i=1}^{N}\frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|} \tag{5}$$

$$\boldsymbol{c} = \frac{\boldsymbol{c\prime}}{\|\boldsymbol{c\prime}\|} \tag{6}$$

non $\boldsymbol{c\prime}$ normalizatu gabeko zentroidea den, $\boldsymbol{x}_i$ aukeratutako generoaren edozein bektore den, N aukeratutako generoko hizlari kopurua den eta $\boldsymbol{c}$ hiperesferan kokatutako zentroidea den. Genero-anbiguoko zentroidea gizon zentroidea eta emakume zentroidea erabiliz kalkula daiteke ondoko formularekin:

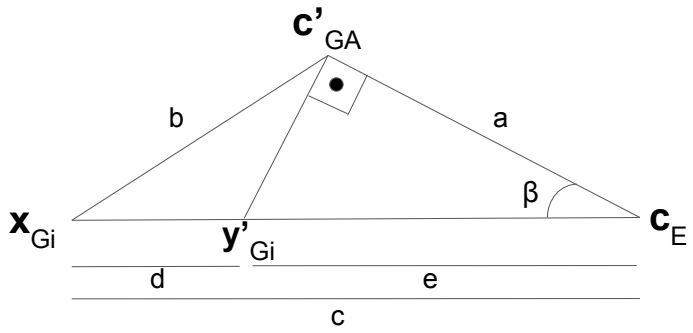
$$\boldsymbol{c\prime}_{GA} = (\frac{\boldsymbol{c}_E - \boldsymbol{c}_G)}{2} \tag{7}$$

Bektore espazio hau hiperesfera baten barruan dagoela kontsideratzen badugu, hizlari baten generoa transformatzeko era egokia beste generoaren zentroidearekiko distantzia angeluarra txikitzea da. MT2 ereduan bezala, transformazioak hizlari-bektorea bi genero zentroideetatik distantzia angeluar berera kokatu behar du. Distantzia euklidearraren txikitzeak distantzia angeluarraren txikitzea dakarrela asumitzen dugunez, MT2 sistemako antzeko operazioa aplikatzen dugu baina kasu honetan transformazioaren norabidea beste generoaren zentroiderantz doa. Soluzio trigonometrikoa 3. Irudian ikus daiteke. Egindako operazioak ondokoak dira:

$$d = c - \frac{2ca^2}{c^2 + a^2 - b^2} \tag{8}$$

$$\boldsymbol{y\prime}_{Gi} = \boldsymbol{x}_{Gi} + d\frac{\boldsymbol{c}_E - \boldsymbol{x}_{Gi}}{\|\boldsymbol{c}_E - \boldsymbol{x}_{Gi}\|} \tag{9}$$

non MT2 sistemako azalpen berak aplika ditzakegun eta aldagaiak 3. Irudian ikus daitezken.



3. irudia. Hizlari-bektore erreal baten transformazioa hiperesferan.

4. Esperimentuak

4.1. Entrenamendu prestaketa

Esperimentuan Google-ek sortutako datu-base hizlari-anitzeko euskarazko zatia erabili dugu entrenamendu datu gisa [13]. Datu-base honek 14 ordu hizketa ditu 52 hizlarirenak, 29 emakume eta 23 gizon, eta hizlari bakoitzaren grabazio kopurua 5 eta 15 minutu artean dago (50-150 esaldi gutxi gorabehera). Testu normalizazio sistema propio bat erabili genuen fonema sekuentziak eta azentu markak lortzeko sarrera testu batetik. Eskuz zuzendu ditugu hainbat izen kanpotarren fonema errepresentazioa ahoskerarekin koherentea izateko. Kaldi [14] ere erabili dugu fonemak grabaketekin lerrokatzeko eta lerrokatze hauek erabili ditugu audioen hasierako eta bukaerako isiluneak ezabatzeko. Datu-basetik 50 esaldi aukeratu ditugu test moduan eta entrenamendutik kendu ditugu esaldi hauek ahoskatzen dituzten grabazio guztiak. Geratzen diren grabazioetatik 10 grabazio hartu ditugu hizlari bakoitzeko balidazio talde bat osatuz, eta gainontzeko grabazio guztiak entrenamendurako erabili ditugu.

MT2 eta MT2-DLN sistemak 400 epoketan entrenatu dira 25eko batch tamainarekin eta balidazio loss onena duen epoka gordez. Adam algoritmoa erabili dugu optimizatzaile bezala 10^{-3} ikasketa-tasarekin. Eredu optimoa 86k pausoren ondoren lortu da MT2 sisteman eta 94k pausoren ondoren MT2-DLN sisteman.

Proposatutako eredu guztiak Waveglow ereduarekin konbinatu dira iragarritako mel espektrogrametatik audio seinaleak lortzeko. Erabilitako Waveglow eredua NVIDIAk partekatutako eredu unibertsalari [15] dohikuntza fina aplikatuz lortu da eta entrenamendu eta balidazio datuak MT2 eta MT2-DLN sistemen berak izan dira. Waveglow eredua 139k pausoz entrenatu dugu 8ko batch tamaina, Adam optimizagailua eta 10^{-4}eko ikasketa-tasa erabiliz.

4.2. Ebaluazioaren prestaketa

Genero-anbiguoko bektoreak bi eratara saiatu gara sortzen: gizon eta emakumeen erdibideko puntua erabiliz, eta antzeko transformazioak aplikatuz ahots errealen hizlari-bektoreari. Transformazioetan, ez dagoenez garbi zein ahots mota den egokiena transformazioarekin genero-anbiguoko ahots bat lortzeko, sei ahots desberdin probatu ditugu: oinarrizko frekuentzia altuena (FA), baxuena (FB), eta erdibidekoa (FE) dituzten gizon (G) eta emakume (E) ahotsak. Ahotsak pYIN metodoaz [16] kalkulatutako F_0 balioen batezbestekoaren bidez aukeratu dira. F_0-aren batezbesteko balioak 1. Taulan erakusten dira.

1. taula. Aukeratutako hizlarien batezbesteko F_0 balioa (Hz).

	FB	**FE**	**FA**
G	97,54	123,83	156,02
E	161,77	206,17	251,56

Transformatutako ahotsak ebaluatzeko, ahots original eta sintetikoekin konparatu ditugu. Sintesian 26 metodo ditugu: 6 hizlari x 2 metodo (MT2, MT2-DLN) x 2 (transformazio gabe, transformatua) + 2 metodo genero bakoitzeko zentroideekin (C). Grabazio errealetan 6 hizlariren grabazioak ditugu. Guztira 32 (26 + 6) hizlari-sistema konbinazio daude, eta ebalutzaile bakoitzari 3 esaldi erakutsi zaizkio konbinazio bakoitzetik ausazko ordenean. Ebaluatutako esaldi bakoitzean, ebaluatzaileak hizlariaren generoa aukeratu behar zuen (gizona edo emakumea) eta 1etik 5erako puntuazioa eman behar zion esaldiaren naturaltasunari.

4.3. Esperimentuaren emaitzak

Genero ebaluazioaren anbiguotasuna neurtzeko Genero-Anbiguotasunaren Puntuazioa (GAP) deitu diogun parametro bat sortu dugu. GAP parametroa ondoko formularekin definitu dugu:

$$GAP = \frac{|| \frac{v_E}{v_E + v_G} - 0,5| - 0,5|}{0,5} \tag{10}$$

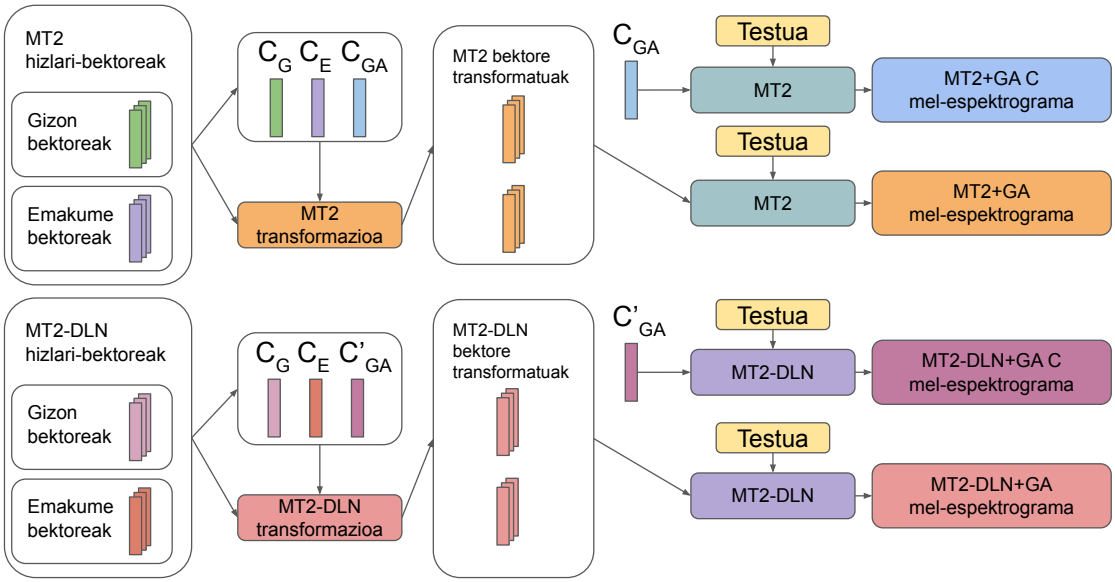
non v_E eta v_G ebaluatzaileek eman dituzten emakume eta gizon botoak diren, eta GAP parametroak 0tik 1era doan puntuazio bat den ebaluatzaileen iritzi dibertsitatea neurtzen duena hizlariaren generoa definitzen dutenean (0 balioak ahobatez genero bat aukeratu dela adierazten duen eta 1 balioak % 50-eko iritzi banaketa perfektua adierazten duen).

Ebaluazioan lortutako naturaltasun balioekin Mean Opinion Score (MOS) balioak kalkulatu dira eta 2. Taulan ikus daitezke % 95-eko konfidantza bitarteekin. Genero klasifikazio subjetiboekin lortutako GAP balioak 3. Taulan ikus daitezke.

Bi emaitza tauletan lerroek hizlaria errepresentatzen dute, genero-anbiguoko zentroidea barne hartuz (C), genero zentroideen batezbesteko bektorearekin sortua. Zutabeek grabazio originalek (GT), MT2 eta MT2-DLN sistemen bidez sortutako audio sintetikoek, eta MT2 eta MT2-DLN sistemei bektoreen genero-anbiguetate manipulazioak aplikatuz lorturiko audioek lortutako puntuazioak (+GA) erakusten dituzte. Genero-anbiguoko ahots bakoitza lortzeko prozesua 4. Irudian ikus daiteke.

2. Taulan ikus dezakegu hainbat genero anbiguoko konbinaziok (G(FA)-MT2-DLN+GA, C-MT2+GA, G(FE)-MT2+GA, E(FB)-MT2+GA) 0,82 puntutik gorako GAP puntuazioa lortzen dutela, beraz ondoriozta dezakegu proposatu ditugun metodoak baliozkoak direla genero-anbiguotasuna duten ahotsak sortzeko. Ezin da ondorio garbirik atera zein ahots mota den egokiena genero-anbiguotasun altua lortzeko, edo zein metodo den onena honetarako ere. E(FA) hizlariaren grabazio originalen anbiguotasun handia ebaluatzaile askok ume bat zela uste zutelako gertatu da.

Ahots transformatuen naturaltasuna igotzen da gizonetan tonua igotzen doan heinean eta emakumeetan tonua jaisten den heinean, honek iradokitzen du tarteko tonuek (tonu altuak gizonetan



4. irudia. Genero-anbiguoko ahotsen transformazio prozesua.

eta tonu baxuak emakumeetan) egokiagoak direla naturaltasuna lortzeko genero-anbiguoko ahotsetan.

Anbiguotasuna eta naturaltasuna kontuan hartuz, G(FA)-MT2-DLN+GA da konbinazio onena: anbiguotasun puntuazio onena lortzen du eta naturaltasun puntuazio altuenetako bat, oso gertu puntuazio altuenetik.

2. taula. Naturaltasunaren MOS balioen emaitzak % 95-eko konfidantza tarteekin.

Hizlaria	GT	MT2	MT2-DLN	MT2+GA	MT2-DLN+GA
G(FB)	4,46 ± 0,16	3,05 ± 0,22	2,93 ± 0,23	2,29 ± 0,17	1,98 ± 0,18
G(FE)	4,22 ± 0,17	3,48 ± 0,18	3,44 ± 0,18	3,00 ± 0,19	2,76 ± 0,20
G(FA)	4,30 ± 0,15	3,60 ± 0,17	3,70 ± 0,19	**3,47 ± 0,19**	3,46 ± 0,17
E(FB)	4,53 ± 0,13	3,53 ± 0,19	3,54 ± 0,18	3,27 ± 0,18	**3,51 ± 0,19**
E(FE)	**4,73 ± 0,12**	**3,95 ± 0,17**	**3,96 ± 0,16**	3,15 ± 0,19	3,35 ± 0,17
E(FA)	4,26 ± 0,15	3,44 ± 0,20	3,34 ± 0,20	2,49 ± 0,20	2,13 ± 0,20
C	—	—	—	3.17 ± 0.17	2,65 ± 0,18

3. taula. GAP balioen emaitzak.

Hizlaria	GT	MT2	MT2-DLN	MT2+GA	MT2-DLN+GA
G(FB)	0,06	0,04	0,00	0,12	0,48
G(FE)	0,06	0,02	0,02	0,88	0,50
G(FA)	0,06	0,22	0,16	0,60	**0,96**
E(FB)	0,20	0,18	0,12	0,84	0,56
E(FE)	0,20	0,14	0,12	0,44	0,52
E(FA)	**0,54**	**0,34**	**0,28**	0,68	0,78
C	—	—	—	**0,92**	0,38

5. Ondorioak eta etorkizunerako lanak

Hizlari-bektoreetan oinarritutako hainbat metodo proposatu ditugu genero-anbiguoko ahotsak sortzeko. Ebaluazioak erakutsi du zenbait jatorrizko ahots eta metodoren konbinazioak emaitza oso onak lortzen dituztela anbiguotasunean, nahiz eta naturaltasunean galerak izan; hala ere, lortutako naturaltasuna onargarria da, transformaziorik gabeko ahots sintetikoen antzeko emaitzak edo hobeak lortzen direlako. Gure ustez, emaitzak onak dira, baina oraindik hobetzeko modukoak, eta ikerketa gehiago egiteko itxaropentsuak.

Etorkizuneko ikerketen artean metodologia hau beste corpus batean probatzea planteatzen dugu. Tarteko tonu balioa duten ahots profesionalak graba genitzake eta Googlen datu-basearei gehitu. Modu honetara hobekuntzak espero genitzake naturaltasunean.

6. Eskerrak

Lan hau Euskarazko laguntzaile adimendun bat garatzeko proiektuaren barruan egin da. Proiektu hau European Language Grid-ek [17] finantzatua izan da bere proiektu pilotoetako bat bezala (Smart euSpeaker proiektua). Eusko Jaurlaritzaren Hazitek programaren (DomEus proiektua) eta Gipuzkoako Probintzia-Kontseiluaren Etorkizuna Eraikiz programaren (Mycroft.eus proiektua) laguntzak ere jaso ditu.

Erreferentziak

[1] WEST, M., KRAUT, R. eta CHEW, H. E., 2019, *I'd blush if I could: closing gender divides in digital skills through education*, Unesco EQUALS, with the support of German Federal Ministry for Economic Cooperation and Development.

[2] TOLMEIJER, S., ZIERAU, N., JANSON, A., WAHDATEHAGH, J. S., LEIMEISTER, J. M. M. eta BERNSTEIN, A., 2021, «Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution», *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Orrialdeak 1–7.

[3] CARPENTER, J., 2019, «Why project q is more than the world's first nonbinary voice for technology», *Interactions*, **26**(6), 56–59.

[4] ŁAŃCUCKI, A., 2021, «Fastpitch: Parallel text-to-speech with pitch prediction», *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orrialdeak 6588–6592, IEEE.

[5] ELIAS, I., ZEN, H., SHEN, J., ZHANG, Y., JIA, Y., SKERRY-RYAN, R. eta WU, Y., 2021, «Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling», *Proc. Interspeech 2021*, Orrialdeak 141–145.

[6] ARIK, S., CHEN, J., PENG, K., PING, W. eta ZHOU, Y., 2018, «Neural voice cloning with a few samples», *Advances in Neural Information Processing Systems*, **31**.

[7] VYLOMOVA, E., RIMELL, L., COHN, T. eta BALDWIN, T., 2015, «Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning», *arXiv preprint arXiv:150901692*.

[8] CHUNG, J. S., HUH, J., MUN, S., LEE, M., HEO, H.-S., CHOE, S., HAM, C., JUNG, S., LEE, B.-J. eta HAN, I., 2020, «In Defence of Metric Learning for Speaker Recognition», *Proc. Interspeech 2020*, Orrialdeak 2977–2981.

[9] JIA, Y., ZHANG, Y., WEISS, R., WANG, Q., SHEN, J., REN, F., NGUYEN, P., PANG, R., LOPEZ MORENO, I., WU, Y. *eta kolaboratzaileak*, 2018, «Transfer learning from speaker verification to multispeaker text-to-speech synthesis», *Advances in neural information processing systems*, **31**.

[10] PRENGER, R., VALLE, R. eta CATANZARO, B., 2019, «Waveglow: A flow-based generative network for speech synthesis», *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orrialdeak 3617–3621, IEEE.

[11] LIU, Z. eta MAK, B., 2020, «Multi-Lingual Multi-Speaker Text-to-Speech Synthesis for Voice Cloning with Online Speaker Enrollment», *Proc. Interspeech 2020*, Orrialdeak 2932–2936.

[12] CAI, W., CHEN, J. eta LI, M., 2018, «Analysis of Length Normalization in End-to-End Speaker Verification System», *Proc. Interspeech 2018*, Orrialdeak 3618–3622.

[13] KJARTANSSON, O., GUTKIN, A., BUTRYNA, A., DEMIRSAHIN, I. eta RIVERA, C. E., 2020, «Open-source high quality speech datasets for basque, catalan and galician», *Proc. of 1st Joint Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL) Workshop (SLTU-CCURL 2020)*, Orrialdeak 21–27, 11–12 May, Marseille, France.

[14] POVEY, D., GHOSHAL, A., BOULIANNE, G., BURGET, L., GLEMBEK, O., GOEL, N., HANNEMANN, M., MOTLICEK, P., QIAN, Y., SCHWARZ, P. *eta kolaboratzaileak*, 2011, «The kaldi speech recognition toolkit», *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF, IEEE Signal Processing Society.

[15] VALLE, R., LI, J., PRENGER, R. eta CATANZARO, B., 2020, «Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens», *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orrialdeak 6189–6193, IEEE.

[16] MAUCH, M. eta DIXON, S., 2014, «pyin: A fundamental frequency estimator using probabilistic threshold distributions», *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, Orrialdeak 659–663, IEEE.

[17] REHM, G., BERGER, M., ELSHOLZ, E., HEGELE, S., KINTZEL, F., MARHEINECKE, K., PIPERIDIS, S., DELIGIANNIS, M., GALANIS, D., GKIRTZOU, K., LABROPOULOU, P., BONTCHEVA, K., JONES, D., ROBERTS, I., HAJIČ, J., HAMRLOVÁ, J., KAČENA, L., CHOUKRI, K., ARRANZ, V., VASIĻJEVS, A., ANVARI, O., LAGZDIŅŠ, A., MEĻŅIKA, J., BACKFRIED, G., DIKICI, E., JANOSIK, M., PRINZ, K., PRINZ, C., STAMPLER, S., THOMAS-ANIOLA, D., GÓMEZ-PÉREZ, J. M., GARCIA SILVA, A., BERRÍO, C., GERMANN, U., RENALS, S. eta KLEJCH, O., 2020, «European language grid: An overview», *Proceedings of the 12th Language Resources and Evaluation Conference*, Orrialdeak 3366–3380, European Language Resources Association, Marseille, France.
URL https://aclanthology.org/2020.lrec-1.413