Ekaia, ale berezia 2025, 113-124
https://doi.org/10.1387/ekaia.26334

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN 0214-9001 – eISSN 2444-3255

Genero aldetik anbiguoa den hizketaren sintesia euskaraz hizlari-bektore manipulazioaren bidez

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

Xabier Sarasola, Ander Corral, Igor Leturia, Iñigo Morcillo*

Orai NLP Teknologiak, Usurbil

LABURPENA: Genero aldetik anbiguoa den ahotsa duten text-to-speech (TTS) sistemek gero eta interes handiagoa pizten dute; besteak beste, laguntzaile birtualetan eta bozgorailu adimendunetan genero-alborapenak eta estereotipoak saihesteko duten ahalmenagatik. Artikulu honetan, ahots-bihurketarako teknika berriak aplikatu dizkiegu ahots-bektoreei, sare neuronaletan oinarrituta dauden eta genero aldetik anbiguoak diren euskarazko TTS sistemak lortzeko. Hizlari-bektoreak hiztun anitzeko Tacotron 2-a entrenatuz lortu ditugu. Hizlari-bektoreen normalizazioa eta eskala-parametro bat erabiltzen duten eta erabiltzen ez duten sistemak konparatu ditugu, baita genero bakoitzeko batez besteko hizlari bektore eta ahots errealen hizlari bektoreen erabilera sistema horietan. Emaitzek frogatzen dute aurkeztutako metodoak baliozkoak direla genero aldetik anbiguoak diren ahotsak lortzeko eta kalitate onargarria dutela baina hobetu daitezkeela.

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). – x.sarasola@orai.eus – https://orcid.org/

Nola aipatu* / *How to cite: Sarasola, Xabier; Corral, Ander; Leturia, Igor; Morcillo, Iñigo (2025). «Genero aldetik anbiguoa den hizketaren sintesia euskaraz hizlari-bektore manipulazioaren bidez». *Ekaia*, ale berezia 2025, 113-124. (https://doi.org/10.1387/ekaia.26334).

Jasotze-data: 2024, maiatzak 17; Onartze-data: 2024, uztailak 22.

ISSN 0214-9001 - eISSN 2444-3255 / © UPV/EHU Press 2025

Lan hau *Creative Commons Aitortu-EzKomertziala-PartekatuBerdin 4.0 Nazioartekoa* lizentzia baten mende dago

1. **SARRERA**

UNESCOk 2019an argitaratu zuen txosten batean [1], laguntzaile birtualetan dagoen sexismoa nabarmendu eta laguntzaile birtual feminizatuak erabiltzearen ondorio kaltegarri ugari aipatzen ziren. Txostenean, 18 gomendio eman zituzten generoari buruzko joera horiek prebenitzeko. 7. gomendioak, zehazki, proposatzen du ez gizonezkoak ez emakumezkoak ez diren laguntzaile birtualen bideragarritasuna aztertzea. Gainera, gerora frogatu da ahots horiek areagotu egiten dutela laguntzaile birtualekiko konfiantza emakumeengan, haiekiko konfiantza oro har kaltetu gabe [2].

Genero aldetik anbiguoa den ahotsa lortzeari buruzko literatura urria da, eta lan gehienek ahotsaren oinarrizko frekuentziak entzuleek hautemandako generoan duen eragina aztertzen dute [2]; hala da, adibidez, Q genero aldetik anbiguoa den lehen ahotsa [3].[1] Lan horietan, entzule ugarirekin ebaluatzen dira ahots natural eta oinarrizko frekuentzia aldatuko ahots asko, oinarrizko frekuentzia edo ahots-ezaugarri anbiguoenak zein diren ikusteko. 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 bakoitzaren grabazio gutxiago behar dituztelako horretarako [4, 5]. Hala ere, ahots-bektoreen manipulazioaren bidez ahots berriak sortzea ez da izan asko landu den ikerketa-lerro bat. Gu Arik eta besteren lanean oinarritu gara [6]; hain zuzen, hitz-bektoreetan aplikatzen diren manipulazioak [7] aplikatzen dituzte hizlari-bektoreetan genero- eta azentu-transformazioak (Britainia Handiko ingelesetik AEBko ingelesera, eta alderantziz) lortzeko. Genero aldetik anbiguoak diren ahotsak lortzeko aurreko saiakerek PCA analisia erabili dute hizlari bektoreen dimentsioa murriztu eta erdibideko genero bektoreak espazio murriztu horretan lortzeko [8, 9]. Gure kasuan, erdibideko genero-eraldaketak hizlari bektoreen dimentsio originaletan egin ditugu. Gainera, metodo berri bat proposatuko dugu hizlari-bektoreak entrenatu eta manipulatzeko, distantzia angeluarreko metodoek hizlarien ezagutzaren artearen egoeran izan duten arrakasta kontuan hartuta [10]. Sortu ditugun ahots sintetikoen adibideak demo webgunean entzun daitezke.[2]

[1] https://genderlessvoice.com

[2] https://orai-nlp.github.io/Genderless-demos

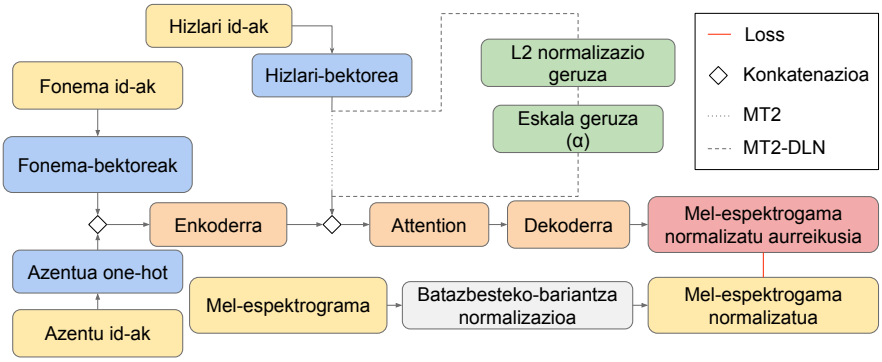
2. HIZLARI ANITZEKO HIZKETAREN SINTESI-EREDUA

Gure hizketa-sintesi sistema ohiko bi etapatan banatuta dago: parametroen sorkuntza eta vocoder-a. Parametro-sorkuntzak hizlari-bektore taula erabiltzen duen Tacotron 2 hizlari anitzeko sistemaren egitura dauka [11], baina aldaketa batzuk gehitu dizkiogu. Horrez gain, genero aldetik anbiguoak diren ahotsak sortzeko, bi teknika probatu ditugu: MT2 eta MT2-DLN. Datozen azpiataletan deskribatuko ditugu. Gure ereduak NVIDIAren Tacotron 2 biltegian oinarrituta daude.[3]

Waveglow erabili dugu vocoder gisa [12], NVIDIAk partekatutako inplementazioa erabiliz.[4]

2.1. MT2

MT2 teknika Tacotron 2 hizlari anitzeko sisteman oinarritzen da [11], zeina hizlari-bektoreek baldintzatuta baitago, baina sarrera- eta irteera-informazioa eta sare neuronalaren galera aldatu ditugu. Horrez gain, fonema-sekuentziak erabili ditugu sarreran, eta, hala, fonema bakoitzean azentuaren presentzia konkatenatu dugu 2-Dko one-hot bektore baten bidez [13]. Irteeran, zero batezbesteko eta unitate-bariantzako normalizazioa aplikatu diogu mel-espektrogramari, entrenamenduko mel-espektrograma erreferentziatzat hartuz. Inferentzian, sarearen irteera denormalizatu dugu, vocoder-ari pasatu aurretik. *L2 loss*-a erabili dugu jatorrizko hizlari bakarreko Tacotron 2-an bezala. Sistemaren eskema orokorra 1. irudian ikus dezakegu.



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

[3] https://github.com/NVIDIA/tacotron2
[4] https://github.com/NVIDIA/waveglow

2.2. **MT2-DLN**

MT2-DLN teknika MT2 sistema bezalakoa da, baina Deep Length Normalization (DLN) aplikatu dugu L2 normalizazio-geruzaren eta ikasgarria den eskala-parametro baten bidez [14]. Horren helburua hizlari-bektore guztiak tamaina ikasgarriko hiperesfera batean kokatzean datza. Sarean erabilitako bektoreen luzera berdina denez, luzera normalizatuaren bertsioan egin diezazkiekegu transformazioak, ondoren berriz sarean erabiltzeko. Horrek esklusiboki distantzia angeluarrekin lan egitea ahalbidetzen digu. Sarearen eskema 1. irudian ikus daiteke, pauso gehigarriak markaturik daudela.

3. **HIZLARI-BEKTOREEN MANIPULAZIOA**

Genero aldetik anbiguoak diren bektoreak kalkulatzeko, zenbait operazio aplikatu dizkiegu entrenatutako bektoreei. Bi mekanismo probatu ditugu:

- Gizonezkoen eta emakumezkoen batez besteko bektoreak kalkulatu eta bien artean dagoen bektorea bilatu.
- Ahots erreal baten bektorea manipulatu, genero bakoitzaren batez besteko bektoreen artean dagoen genero-anbiguotasuna espaziora mugitzeko.

Proposatzen dugun sistema bakoitzean, ezberdinak dira operazio horiek, MT2an distantzia euklidearrak eta MT2-DLNan distantzia angeluarrak hartu baititugu kontuan. Eragiketa bakoitza hurrengo azpiataletan azalduko dugu.

3.1. **MT2 bektore manipulazioa**

MT2 sisteman, [6]-ko logika euklidearrari jarraitu diogu bektore berriak lortzeko. Hala, genero bakoitzaren batez besteko bektorea kalkulatu dugu genero bakoitzaren zentroide gisa, formula hau erabiliz:

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

non $\boldsymbol{x}_i$ aukeratutako generoko edozein hizlariren bektorea den eta N genero horretako hizlari kopurua. Ondoren, genero aldetik anbiguoa den zentroidea ($\boldsymbol{c}_{GA}$) kalkulatu dugu, modu honetan:

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

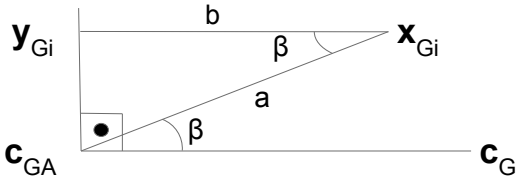
non $\boldsymbol{c}_G$ gizon zentroidea den eta $\boldsymbol{c}_E$ emakume zentroidea den.

Edozein hizlari-bektore genero aldetik anbiguoa den bektore bihurtzeko, transformazio-bektorearen norabidea eta modulua kalkulatu behar ditugu. 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 hori genero aldetik anbiguoa den zentroidearen eta genero-zentroideen arteko norabidearen ortogonala den planoan kokatuta dago. Hizlari-bektorea plano horretara eramaten duen modulua trigonometria aplikatuz (ikus 2. irudia) lor dezakegu, honako eragiketa honen bidez:

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

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

non $\boldsymbol{y}_{Gi}$ hizlari-bektore bihurtua den eta $\boldsymbol{x}_{Gi}$ gizon-bektore originala. Gainontzeko aldagaiak 2. irudian definituak daude. Emakume hizlariaren bektorearen bihurketan ($\boldsymbol{xEi}$), genero-marka guztiak alderantzikatu ditugu formulan. Inferentzia egiteko garaian, genero aldetik anbiguoa den bektorea eta transformatutako bektoreak erabili ditugu genero aldetik anbiguoak diren hizlarien mel-espektrogramen sintesia egiteko.



2. irudia. Transformazioa euklidear espazioan.

3.2. **MT2-DLNko bektoreen manipulazioa**

MT2-DLN sisteman, bektoreak normalizatu ditugu sarean sartu aurretik. Horrek esan nahi du entrenatutako bektore guztiak normaliza ditzakegula eta unitate luzerako hiperesfera batean egin ditzakegula transformazio guztiak. Horrek aukera ematen digu transformazioak distantzia angeluarretan egiteko. Luzera normalizatua daukaten zenbait bektoreren batez besteko angelua era honetan lor daiteke:

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

$$\boldsymbol{c} = \frac{\boldsymbol{c}'}{\left\|\boldsymbol{c}'\right\|} \tag{6}$$

non $\boldsymbol{c}'$ 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 aldetik anbiguoa den zentroidea gizon zentroidea eta emakume zentroidea erabiliz kalkula daiteke formula honekin:

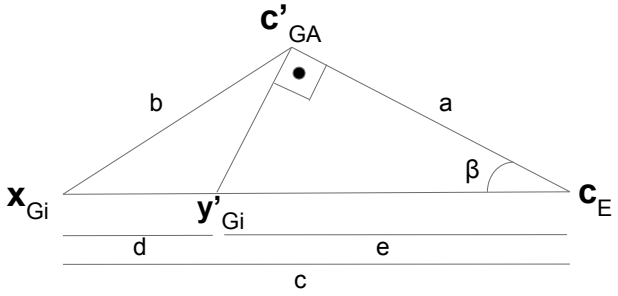
$$\boldsymbol{c}'_{GA} = \frac{(\boldsymbol{c}_E - \boldsymbol{c}_G)}{2} \tag{7}$$

Bektore-espazio hori hiperesfera baten barruan dagoela kontuan hartzen badugu, hizlari baten generoa transformatzeko era egokia beste generoaren zentroidearekiko distantzia angeluarra txikitzea da. MT2 ereduan bezala, transformazioak bi genero zentroideetatik distantzia angeluar berera kokatu behar du hizlari-bektorea. Distantzia euklidearra txikitzeak distantzia angeluarra txikitzea dakarrela onartzen dugunez, MT2 sistemako antzeko operazioa aplikatuko dugu, baina, kasu honetan, transformazioaren norabidea beste generoaren zentroiderantz doa. Soluzio trigonometrikoa 3. irudian ikus daiteke. Honako eragiketa hauek egin ditugu:

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

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

non MT2 sistemako azalpen berak aplika ditzakegun eta aldagaiak 3. irudian ikus daitezkeen.



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

4. **ESPERIMENTUAK**

4.1. **Entrenamenduaren prestaketa**

Esperimentuan, Googlek sortutako hizlari anitzeko datu-baseko euskarazko zatia erabili dugu entrenamendurako datu gisa [15]. Datu-base horrek 52 hizlariren 14 hizketa-ordu biltzen ditu (29 emakume eta 23 gizon), eta hizlari bakoitzaren grabazioaren denbora 5 eta 15 minutu artekoa da (50-150 esaldi gutxi gorabehera). Testu-normalizaziorako sistema propio bat erabili dugu fonema-sekuentziak eta azentu-markak sarreratestu batetik lortzeko. Eskuz zuzendu ditugu hainbat izen, atzerriko fonemen errepresentazioa ahoskerarekin koherentea izan dadin. Kaldi [16] ere erabili dugu, fonemak grabazioekin lerrokatzeko, eta lerrokatze horiek baliatu ditugu audioen hasierako eta bukaerako isiluneak ezabatzeko. Datu-basetik 50 esaldi aukeratu ditugu, test moduan, eta entrenamendutik kendu ditugu esaldi horiek ahoskatzen dituzten grabazio guztiak. Gainontzeko grabazioetatik 10 hartu, eta hizlari bakoitzeko baliozkotze talde bat osatu dugu. Gainontzeko grabazio guztiak entrenamendurako erabili ditugu.

MT2 eta MT2-DLN sistemak 400 epochetan entrenatu ditugu, 25eko batch tamainarekin eta baliozkotze loss onena duen epocha gordez. Adam algoritmoa erabili dugu optimizatzaile gisa, 10^{-3} ikasketa-tasarekin. Eredu optimoa 86k pausoren ondoren lortu dugu MT2 sisteman, eta 94k pausoren ondoren MT2-DLN sisteman.

Proposatutako eredu guztiak Waveglow ereduarekin konbinatu ditugu, iragarritako mel espektrogrametatik audio-seinaleak lortzeko. Erabilitako Waveglow eredua NVIDIAk partekatutako eredu unibertsalari [17] doikuntza fin bat aplikatuz lortu dugu, eta entrenamendu- eta baliozkotze-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 aldetik anbiguoak diren bektoreak bi eratara saiatu gara sortzen: gizon eta emakumeen erdibideko puntua erabiliz, eta antzeko transformazioak aplikatuz ahots errealen hizlari-bektoreari. Transformazioetan, ez dagoenez garbi zer ahots mota den egokiena transformazioarekin genero aldetik anbiguoa den ahots bat lortzeko, sei ahots probatu ditugu: oinarrizko frekuentziarik altuena (FA), baxuena (FB) eta erdibidekoa (FE) dituzten gizon (G) eta emakume (E) ahotsak. Ahotsak pYIN metodoaz [18] 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 × 2 metodo (MT2, MT2-DLN) × 2 (transformaziorik gabe, transformatua) + 2 metodo genero bakoitzeko zentroideekin (C). Grabazio errealetan, 6 hizlariren grabazioak ditugu. Guztira, 32 (26 + 6) hizlari-sistema konbinazio daude, eta ebaluatzaile bakoitzari 3 esaldi erakutsi dizkiogu konbinazio bakoitzetik, ausazko ordenan. 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**

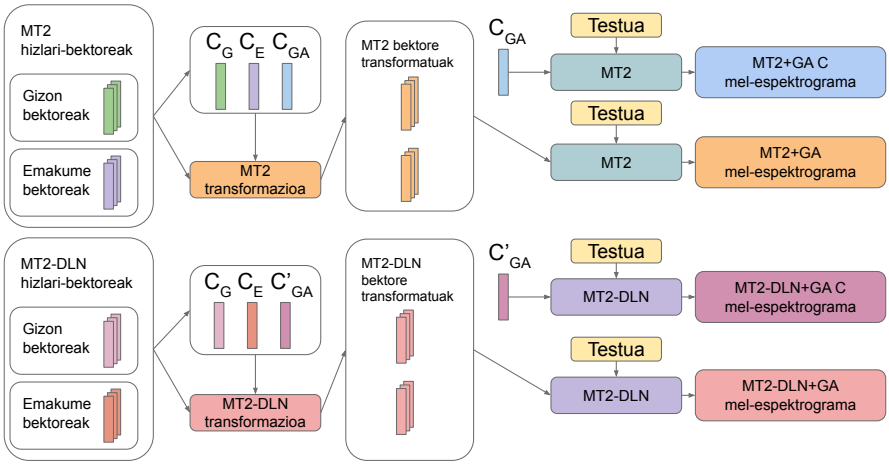
Genero-anbiguotasuna neurtzeko Genero Anbiguotasunaren Puntuazioa (GAP) deitu diogun parametro bat sortu dugu. GAP parametroa ondoko formularekin definitu dugu:

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

non v_E eta v_G ebaluatzaileek gizon eta emakumeei eman dizkieten botoak diren, eta GAP parametroa 0tik 1era doan puntuazio bat den, ebaluatzaileen iritzi-aniztasuna neurtzen duena hizlariaren generoa definitzen dutenean (0 balioak adierazten du genero bat aho batez aukeratu dela, eta 1 balioak, berriz, % 50eko iritzi-banaketa perfektua).

Ebaluazioan lortutako naturaltasun-balioekin, Mean Opinion Score (MOS) balioak kalkulatu ditugu, eta 2. taulan ikus daitezke % 95eko konfiantza-tarteekin. Genero-sailkapen subjektiboekin lortutako GAP balioak 3. taulan ikus daitezke.

Bi emaitza-tauletan, lerroek hizlaria ordezkatzen dute, genero aldetik anbiguoa den zentroidea barne hartuz (C), zeina genero zentroideen batezbesteko bektorearekin sortua baita. Zutabeetan adierazi ditugu grabazio originalen (GT), MT2 eta MT2-DLN sistemen bidez sortutako audio sintetikoen, eta MT2 eta MT2-DLN sistemei bektoreen genero-anbiguotasunaren manipulazioak aplikatuz lorturiko audioen puntuazioak (+GA). Genero aldetik anbiguoa den ahots bakoitza lortzeko prozesua 4. irudian ikus daiteke.



4. irudia. Genero-anbiguoko ahotsen transformazio prozesua.

2. taulan ikus dezakegu genero aldetik anbiguoak diren zenbait konbinaziok (G(FA) – MT2 – DLN + GA, C – MT2 + GA, G(FE) – MT2 + GA, E(FB) – MT2 + GA) 0,82 puntutik gorako GAP puntuazioa lortu dutela; beraz, ondoriozta dezakegu proposatu ditugun metodoak baliozkoak direla genero-anbiguotasuna duten ahotsak sortzeko. Ezin da ondorio garbirik atera zer ahots mota den egokiena edo onena genero-anbiguotasun altua lortzeko. E(FA) hizlariaren grabazio originaletan, anbiguotasun handia hauteman dute ebaluatzaile askok, ume bat zela uste zutelako.

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

Gizonen tonua igo eta emakumeena jaisten den heinean, ahots transformatuen naturaltasuna igotzen da, eta horrek iradokitzen du tarteko tonuak (tonu altuak gizonetan eta tonu baxuak emakumeetan) egokiagoak direla genero aldetik anbiguoak diren ahotsetan naturaltasuna lortzeko.

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

5. **ONDORIOAK ETA ETORKIZUNERAKO LANAK**

Hizlari-bektoreetan oinarritutako zenbait metodo proposatu ditugu genero aldetik anbiguoak diren ahotsak sortzeko. Ebaluazioak erakutsi du jatorrizko zenbait ahots eta metodoren konbinazioak emaitza oso onak lortzen dituztela anbiguotasunari dagokionez, 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.

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

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

6. **ESKERRAK**

Lan hau euskarazko laguntzaile adimendun bat garatzeko proiektuaren barruan egin da. Proiektu hau European Language Grid-ek [19] finantzatu du, proiektu pilotu gisa (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] M. WEST, R. KRAUT eta H. E. CHEW. 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] S. TOLMEIJER, N. ZIERAU, A. JANSON, J. S. WAHDATEHAGH, J. M. M. LEIMEISTER eta A. BERNSTEIN. 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] J. CARPENTER. 2019. «Why project q is more than the world's first nonbinary voice for technology», Interactions, **26**(6), 56-59.

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

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

[7] E. VYLOMOVA, L. RIMELL, T. COHN eta T. BALDWIN, 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] K. MARKOPOULOS, G. MANIATI, G. VAMVOUKAKIS, N. ELLINAS, G. VARDAXOGLOU, P. KAKOULIDIS, J. OH, G. JHO, I. HWANG, A. CHALAMANDARIS, P. TSIAKOULIS eta S. RAPTIS, 2023, «Generating Multilingual Gender-Ambiguous Text-to-Speech Voices», *Proc. INTERSPEECH 2023*, orrialdeak 621-625.

[9] C. YU, C. FU, R. CHEN eta A. TAPUS, 2022, «First attempt of gender-free speech style transfer for genderless robot», *2022 17th ACM/IEEE International Conference on Human- Robot Interaction (HRI)*, Orrialdeak 1110-1113, IEEE.

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

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

[12] R. PRENGER, R. VALLE eta B. CATANZARO, 2019, «Waveglow: A flow-based generative network for speech synthesis», *ICASSP 2019-2019 IEEE Inter-*

national Conference on Acoustics, Speech and Signal Processing (ICASSP), Orrialdeak 3617-3621, IEEE.

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

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

[15] O. KJARTANSSON, A. GUTKIN, A. BUTRYNA, I. DEMIRSAHIN eta C. E. RIVERA, 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.

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

[17] R. VALLE, J. LI, R. PRENGER eta B. CATANZARO, 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.

[18] M. MAUCH eta S. DIXON, 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.

[19] G. REHM, M. BERGER, E. ELSHOLZ, S. HEGELE, F. KINTZEL, K. MARHEINECKE, S. PIPERIDIS, M. DELIGIANNIS, D. GALANIS, K. GKIRTZOU, P. LABROPOULOU, K. BONTCHEVA, D. JONES, I. ROBERTS, J. HAJIČ, J. HAMRLOVÁ, L. KAČENA, K. CHOUKRI, V. ARRANZ, A. VASIL, JEVS, O. ANVARI, A. LAGZDIN, Š, J. MEL, N, IKA, G. BACKFRIED, E. DIKICI, M. JANOSIK, K. PRINZ, C. PRINZ, S. STAMPLER, D. THOMAS-ANIOLA, J. M. GÓMEZ-PÉREZ, A. GARCIA SILVA, C. BERRÍO, U. GERMANN, S. RENALS eta O. KLEJCH. 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