VI. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

Aitortu-PartekatuBerdin 4.0

INGENIARITZA ETA ARKITEKTURA

HiTZketan. Ahots-ahots itzulpen elebiduna ikasketa sakona erabiliz

Aitor Bellanco, Ibon Saratxaga, Inma Hernáez, Aitor Soroa, Eva Navas, Gorka Labaka, Xabier de Zuazo, Ander Arriandiaga, Victor García, Jon Sanchez, Christoforos Souganidis, Eneko Aguirre, Asier Herranz, Rodrigo Agerri, German Rigau eta *Gemma Meseguer*

243-250 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.30

ANTOLATZAILEA
BABESLEAK
LAGUNTZAILEAK

HiTZketan. Ahots-ahots itzulpen elebiduna ikasketa sakona erabiliz

Aitor Bellanco[1], Ibon Saratxaga[1], Inma Hern´aez[1], Aitor Soroa[2], Eva Navas[1], Gorka Labaka[2], Xabier de Zuazo[1], Ander Arriandiaga[1] Victor García[1], Jon Sanchez[1], Christophoros Souganidis[1], Eneko Agirre[2], Asier Herranz[1], Rodrigo Agerri[2], German Rigau[2], Gemma Meseguer[1],

[1] HiTZ Zentroa - Aholab, Euskal Herriko Unibertsitatea (UPV/EHU)
[2] HiTZ Zentroa - Ixa, Euskal Herriko Unibertsitatea (UPV/EHU)

aitor.bellanco@ehu.eus

Laburpena

HiTZketan proiektuak ahotsetik ahotserako (S2S) itzulpen automatikoko sistema berritzaile bat garatu du, euskararen eta gaztelaniaren arteko komunikazio arina ahalbidetzen duena. Sistema horren erabiltzaileak bi hizkuntza horietako edozeinetan hitz egin dezake, eta sistema bera arduratzen da diskurtsoaren edukia kontrako hizkuntzara itzultzeaz. Gainera, komunikazioaren esperientzia eta naturaltasuna hobetzeko, sistemak testu-itzulpena sortzeaz gain, audio bihurtzen du, jatorrizko hiztunaren intonazioa eta ahots-ezaugarriak imitatzen dituen ahots pertsonalizatua erabiliz.

Hitz gakoak: Ahots-arteko itzulpena, Hizketa-ezagutza automatikoa (ASR), Itzulpen automatiko neuronala (MT), Ahots-sintesi pertsonalizatua (p-TTS), Zero-adibide ikaskuntza.

Abstract

HiTZketan has developed an innovative speech-to-speech (S2S) automatic translation system that enables seamless communication between Basque and Spanish. The system allows users to speak naturally in either language while instantly translating their speech into the other. To enhance the experience and ensure more natural communication, it not only provides a text translation but also converts it into audio, using a customized voice that replicates the speaker's intonation and vocal characteristics.

Keywords: *Speech-to-speech translation, Automatic speech recognition (ASR), Neural machine translation (MT), Personalized speech synthesis (p-TTS), Zero-shot learning.*

1 Sarrera eta motibazioa

Gero eta globalizatuagoa den mundu honetan, funtsezkoa da hizkuntza ezberdinak hitz egiten dituzten pertsonen arteko komunikazioa erraztea. Hala ere, interakzio hori zailtzen duten hizkuntza-oztopoak daude, batez ere hiztun anitzak ez duten hizkuntzetan edo baliabide mugatuak dituzten testuinguruetan, hala nola euskaran. Euskara hizkuntza gutxitua da, eta hiztun-komunitate txikia du gaztelaniarekin alderatuta (Hualde eta Ortiz de Urbina, 2011).

Euskararen erabilera zaintzeko eta sustatzeko ahaleginak burutu diren arren, euskararen eta gaztelaniaren arteko komunikazioa errazteko tresna teknologiko aurreratuen falta erronka iraunkorra izan da. Euskara bezalako hizkuntza gutxituetan, erronkak are handiagoak dira, baliabide linguistikoak eta entrenamendu-datuak mugatuak direlako.[1] Dauden itzulpen automatikoko sistemak hizkuntza maioritarioetan zentratu ohi dira, eta, kasu askotan, ez dute itzulpen arin eta naturalik eskaintzen egitura gramatikal eta fonetiko oso desberdinak dituzten hizkuntzen artean, hala nola euskara eta gaztelania.[2]

Testuaren itzulpen automatikoaren erabilera zabalduta dago gizartean, baina S2S sistema urrats bat harago doa, eta hizkuntza batetik bestera ahotsa itzultzea ahalbidetzen du, hiztunaren nortasuna zaintzen duen bitartean.

Ahotsetik ahotserako itzulpen-sistema gehienek ez dituzte kontuan hartzen jatorrizko hiztunaren ahots-

[1] https://www.euskadi.eus/contenidos/informacion/gaitu_plana/es_def/adjuntos/GAITU_gaz.pdf
[2] https://www.langune.eus/es/noticias/modela-el-traductor-automatico-castellano-euskera-desarrollado-con-inteligencia-artificial-ya-disponible

ezaugarriak, eta horrek komunikazioaren naturaltasuna eta eraginkortasuna murrizten ditu. HiTZketan proiektuan sistema berritzaile bat garatu da, euskararen eta gaztelaniaren arteko itzulpen automatikoa egiteko aukera ematen duena, bi hizkuntza horietako hiztunen arteko komunikazio arin eta naturala erraztuz.

Sistemaren ezaugarri nabarmenetako bat jatorrizko hiztunaren intonazioa eta ahots-ezaugarriak imitatzen dituen ahots sintetizatu bat sortzeko duen gaitasuna da, komunikazio-esperientzia hobetzeko. Sistemak teknologia aurreratuak integratzen ditu, hala nola hizketaren ezagutza automatikoa (Automatic Speech Recognition, ASR), itzulpen automatiko neuronala (Machine Translation, MT) ahots pertsonalizatuaren sintesia (Personalized Text To Speech, p-TTS), eta zero-shot learning teknikak baliatzen ditu datu mugatuak dituzten agertokietan eraginkortasuna hobetzeko.

2 Arloko egoera eta ikerketaren helburuak

2.1 Arloko Egoera

Ahotsetik ahotserako itzulpen automatikoa (Speech-to-Speech Translation, S2S) ahotsaren prozesamenduaren eta hizkuntza naturalaren prozesamenduaren (Natural Language Processing, NLP) barruan etengabe eboluzionatzen ari den ikerketa-eremua da. Testuaren itzulpen automatikoko sistemek zehaztasun eta abiadura handia lortu duten arren, ahotsetik ahotserako itzulpenak erronka gehigarriak ditu, teknologia ugari integratu behar direlako, hala nola hizketaren ezagutza automatikoa (ASR), itzulpen automatikoa (MT) eta ahots-sintesia (TTS). Sistema horiek jardun behar dute komunikazioaren naturaltasuna mantenduz eta, kasu batzuetan, ahots sintetizatua pertsonalizatu behar dute, jatorrizko hiztunaren ahotsarekin bat etor dadin.

Hizketa-teknologiak euskarara eta baliabide gutxiko beste hizkuntzetara egokitzeko ikerketa-proiektuak ugaritu dira, aurrerapen nabarmenak lortuz ahotsaren ezagutzan eta itzulpen automatikoan. Hala ere, teknologia horiek ahotsetik ahotserako itzulpen-sistema batean integratzea, eraginkortasunez funtzionatuko duena, ikerketa-arlo aktiboa da oraindik ere, eta inpaktu-potentzial handia du.

2.2 Ikerketaren helburuak

HiTZ Hizkuntza Teknologiako Euskal Zentroak garatu duen HiTZketan proiektuaren helburu nagusia euskararako ahotsetik ahotserako itzulpen automatikoko sistema integratu bat sortzea da, haren eragina ebaluatuz eta euskal komunitatean hedatzea web plataforma erakusle baten bidez. Zerrenda honetan aurkezten dira Ikerketaren helburu espezifikoak:

1. **Ahotsetik ahotsera itzulpen sistema integratua garatzea**
   Hizketa automatikoki ezagutzeko teknologia aurreratuak (ASR), itzulpen automatiko neuronala (MT) eta ahots-sintesi pertsonalizatua (p-TTS) konbinatzen dituen sistema bat sortzea, euskararen eta gaztelaniaren arteko komunikazio arin eta naturala ahalbidetzeko.
2. **Euskararako ASR ereduen hobekutza**
   Euskarazko ahotsa hobeto ezagutzea, hizkuntza horretako datu espezifikoak dituzten neurona-sare sakonen ereduak entrenatuz.
3. **Lagin bakarreko ahots pertsonalizatuen sintesizagailuak sortzea**
   Jatorrizko hiztunaren ahots-ezaugarriak imitatuko dituen ahots sintetizatu bat sortzea, zero adibide ikasteko teknikak erabiliz (zero-shot learning), komunikazio-esperientzia pertsonalizatzeko. Horretarako ahotsa sintetizatzeko dauden corpusak moldatu dira, sistema bi hizkuntzetan eraginkorra izango dela bermatzeko.
4. **Euskara-gaztelera itzulpen sistema automatikoaren hobekuntza**
   Datu paraleloen eta sintetikoen bolumen handiak erabiltzea itzulpen automatikoko ereduak prestatzeko, euskararen eta gaztelaniaren arteko itzulpenen kalitatea eta abiadura hobetuz.
5. **Kapitalizazio eta puntuazio modulu baten garapena** Puntuazio- eta kapitalizazio-modulu bat garatzea, testuaren egitura berrezartzeaz gain, ASRren akatsak zuzendu eta irakurgarritasuna hobetzeko.
6. **Web erakusle bat sortzea**
   Web-interfaze erabilerraza garatzea, erabiltzaileek ahotsetik ahotserako itzulpen-sistema probatzeko aukera izan dezaten. Erakuslea exekutatzeko GPUak dituen zerbitzari-azpiegitura bat instalatuko da, adimen artifizialeko ereduak eraginkortasunez gauzatzeko.

7. **Kultura- eta gizarte-inklusioa sustatzea**
   Euskararen eta gaztelaniaren hiztun-komunitateen arteko komunikazioa erraztea. Erakustea nola teknologia aurreratuen konbinazioak arazo konplexuak konpon ditzakeen komunikazio eleanitzaren esparruan.

3 Ikerketaren muina

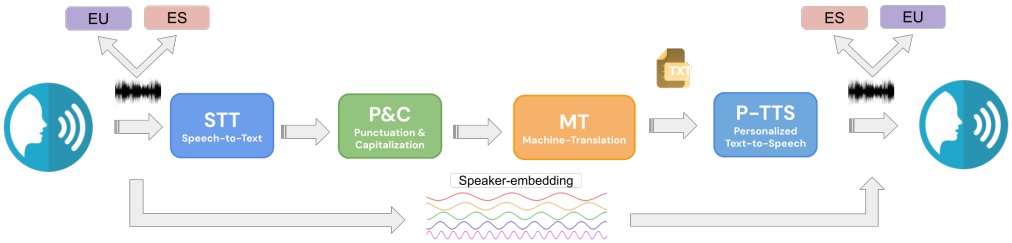
Atal honetan, HiTZketan proiektuaren nukleo tekniko eta metodologikoari eusten dioten funtsezko elementuak aurkezten dira. Sistemaren osagaien taulak, irudiak eta deskribapen zehatzak jasotzen dira, bai eta proiektuaren garapenean lortutako metrikak eta emaitza garrantzitsuak ere.

3.1 HiTZketan sistemaren arkitektura orokorra

Sistemaren arkitekturak ahozko prozesamenduaren fluxua islatzen du bi hizkuntzatan (euskaraz eta gaztelaniaz). Lehenik, sistemak euskarazko (EU) edo gaztelaniazko (ES) ahots-sarrera bat jasotzen du, eta testu bihurtzen du ahotsa ezagutzeko eredu baten bidez (STT). Gero, puntuazio- eta kapitalizazio-zuzenketak aplikatzen dira (P & C - Punctuation & Capitalization) testua beste hizkuntzara itzuli aurretik. Itzulpena transformer-etan oinarritutako eredu bat erabiliz (MT) egiten da. Azkenik, itzulitako testua ahots bihurtzen da ahots-sintesi pertsonalizatuko sistema baten bidez (P-TTS), zeinak jatorrizko hiztunaren ezaugarriei eusten dien ”speaker embedding” eredu baten bidez. Emaitza itzulitako hizkuntzako audio bat da (ES edo EU), ahotsetik ahotsera bihurtzeko zikloa osatuz.

1. irudian, HiTZketan sistemako datu-fluxua erakusten duen diagrama ikus daiteke, eta hurrengo azpiataletan osagai guztien deskribapena egiten da.

1. Irudia: HiTZketan sistemaren arkitektura-diagrama



3.2 Automatic Speech Recognition (ASR)

3.2.1 Ereduak

Atal honetan, sisteman ezarritako hiru ASR ereduak jasotzen dira. Nvidiak garatutako Conformer-Transducer arkitekturan oinarritzen den lehenengo eredua[3] , gaztelaniazko audioa transkribatzeko diseinatuta dago. Bigarren eredua berriz, euskarazko transkripziorako optimizatuta dago. HiTZ zentroak garatu du eta, aurrekoa bezala, Conformer-Transducer arkitekturan oinarrituta dago[4]. Azkenengo eredua, Whisperren eredu ertaina oinarri hartuta[5] moldatu da euskararako. Lehenengo bi ereduak Nvidia-ren NeMo toolkitarekin entrenatu dira. Aldiz, Whisper Medium ereduak transformerren arkitektura darabil, eta ikuspegi eleanitzarekin moldatuta (finetuned) izan da. Hori dela eta, bereziki egokia da transkripzioa hizkuntza ugaritan egiteko, euskara eta gaztelania barne.

3.2.2 Datu multzoak

Ahotsa ezagutzeko sistemak entrenatzeko eta ebaluatzeko, hainbat datu-multzo erabili dira. Mozillak garatutako Common Voice datu-multzoa hizkuntza anitzeko corpus bat da, eta borondatezko ahotsen grabazioak biltzen ditu hainbat hizkuntzetan, horien artean euskara. OpenSLR (SLR76), (**?**), kalitate handiko audio transkribatuak eskaintzen ditu, boluntarioek grabatutako euskarazko esaldiekin. Gainera, Multilingual LibriSpeech (MLS) corpus eleanitz zabala da, LibriVoxen audioliburuetatik eratorria, eta hizkuntza ugari biltzen ditu. Azkenik, Basque Parliament corpusak Eusko Legebiltzarreko saioen grabazioak ditu, eta baliabide baliotsua eskaintzen du euskarazko testuinguru politiko eta formaletan ahotsaren aintzatespena ebaluatzeko.

[3] https://huggingface.co/nvidia/stt_es_conformer_transducer_large
[4] https://huggingface.co/HiTZ/stt_eu_conformer_transducer_large
[5] https://huggingface.co/zuazo/whisper-medium-eu-cv16_1

3.2.3 ASR ereduen ebaluazioa

Ahots-ezagutza sistema baten zehaztasuna neurtzeko WER (Word Error Rate) metrika erabili da. 1. taulan, aipatutako hiru ereduen emaitzak aurkezten dira:

1. Taula: ASR ereduen ebaluazioaren emaitzak hainbat datu-multzotan.

Ereduak	Ebaluazio datu-multzoa	WER
nvidia/stt_es_conformer_transducer_large	Common Voice 7-0-6 es	5,2%
	Multilingual LibriSpeech es	3,2%
HiTZ/stt_eu_conformer_transducer_large	Mozilla Common Voice 16.1 eu	2,79%
	Basque Parliament eu	3,83%
zuazo/whisper-medium-eu-cv16_1	Common Voice 13.0 eu	14,02%
	SLR76 eu	4,92%
	Common Voice 13.0 es	23,2%
	MLS es	5,33%

Oro har, ereduek jarduera sendoa erakusten dute, batez ere Conformer-Transducer arkitekturakoek, gaztelerazko Conformer-Transducer ereduak %5,2ko WERa lortzen baitu Common Voice 7-0-6an eta %3,2koa Multilingual LibriSpeechen. Euskararentzat (eu), Conformer-Transducer ereduak %2,79ko WERa lortzen du Mozilla Common Voice 16.1 eremuan, eta horrek zehaztasun maila handia iradokitzen du. Aldiz, whisper ereduak errendimendu aldakorragoa erakusten du, eta WERa nabarmen altua da SLR76 eu-n (%23,2) eta Common Voice 13.0 eu-n (%14,02). Horrek esan nahi du zailtasunak daudela euskarazko zenbait datu-multzo transkribatzeko. Konparazio bat eginez, bere espainierazko errendimendua sendoagoa da, Common Voice 13.0 es %4,92ko WERarekin. Emaitza horiek eredua hizkuntzara eta datuen berariazko menderatzera egokitzearen garrantzia nabarmentzen dute (de Zuazo *et al.*, 2025).

3.3 Itzulpen Automatiko Neuronala (MT)

3.3.1 Ereduak

Ezarritako Machine Translation moduluak bi etapako ikuspegia jarraitzen du itzulpenaren kalitatea hobetzeko. Lehenik, puntuazio- eta kapitalizazio-eredu bat erabiltzen da, sarrerako testuaren egitura gramatikala berrezartzeko, letra larriak eta puntuazio-zeinu egokiak gehituz. Hori funtsezkoa da hurrengo fasean edukia behar bezala interpretatuko dela bermatzeko. Ondoren, testu prozesatua MarianMT-n[6] sartzen da, transformerretan oinarritutako itzulpen eredua, euskarazko itzulpen zehatzak eta eraginkorrak eskaintzeko optimizatua.

3.3.2 Emaitzen analisia

Machine Translation-en eredu bat ebaluatzeko, BLEU (Bilingual Evaluation Understudy)[7] metrika erabiltzen da. Metrika honek sortutako itzulpenen kalitatea neurtzeko aukera ematen du, horretarako, testu itzuliaren eta erreferentziaren artean n-gramen berdintasunean oinarritutako zehaztasuna kalkulatzen du, itzulpen laburregiak zigortuz eta jatorrizko edukiarekiko arintasuna eta fideltasuna mantentzen dituztenak lagunduz.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

2. taulan , itzulpen-ereduen ebaluazio-metrikak alderatzen dira, erabilitako sarrera-motaren arabera. Alde batetik, kapitalizazioa eta puntuazioa (C&P) duen sarrerako testuarekin (orig) lan egiten duten ereduek emaitzarik onenak lortzen dituzte BLEU metrikaren arabera. Aldiz, testu lauak (C&Prik gabe) (norm) erabiltzen dituztenek emaitza txarragoak eskuratzen dituzte. Hala ere, testu normalizatuei itzulpenaren aurretik puntuazio- eta kapitalizazio-eredu bat (C&PR) aplikatzen zaienean, emaitzak nabarmen hobetzen dira, jatorrizko testuekin lortutakoetara hurbilduz.

[6] https://huggingface.co/docs/transformers/model_doc/marian
[7] https://es.wikipedia.org/wiki/BLEU

2. Taula: Itzulpen-ereduen ebaluazio metriken konparaketa.

Sarrera	Hizkuntza itzulketa	BLEU
es_orig	es -> eu	13,7
es_norm	es -> eu	8,3
es_norm	es_norm -> es (C&PR) -> eu	13,0
es_norm	es_norm -> eu	12,6
eu_orig	eu -> es	19,7
eu_norm	eu -> es	12,8
eu_norm	eu_norm -> eu (C&PR) -> es	18,4
eu_norm	eu_norm -> es	18,6

3.4 Ahots pertsonalizatuaren sintesia (p-TTS)

Sistema honetarako, Ahots Pertsonalizatuko Testu-Hizkuntza Sistema (p-TTS) bat garatu da, euskaraz eta gaztelaniaz ahots sintetikoak sortzeko ikuspegi eleanitza erabiltzen duena. Helburu nagusia ahots naturala sortzea da, baina baita jatorrizko hiztunaren ezaugarri bereziei eustea ere, baita beste hizkuntza batean sortzen denean ere. Jarraian, erabilitako teknologia eta haren eraginkortasuna bermatzen duten funtsezko metrikak deskribatzen dira.

3.4.1 Eredua

P-TTS sistemak ahots-sintesi arkitektura aurreratua erabiltzen du, YourTTS (Casanova *et al.*, 2022b)-n oinarritua, eta ikasketa sakoneko teknikak eta bariazio-inferentzia konbinatzen ditu ahots pertsonalizatuak sortzeko. Eredu honek hizkuntza anitzeko egokitzapena eta ahots-klonazioa ahalbidetzen ditu, eta horrek esan nahi du hizkuntza desberdinetan hitz egiteko gaitasuna duela., jatorrizko hiztunaren ahots-ezaugarriak gordetzen dituen bitartean. Proiektu honetan ingelera, portugalera, frantsesa, euskara, gaztelania, katalana eta galizierarekin entrenatutako eredu bat erabili da. Eredu hau Aholab taldean garatu zen aurreko lan baten barruan.

3.4.2 Datu-multzoak

Atal honek ahots-sintesiko eredu eleaniztuna (TTS) prestatzeko erabiltzen diren datu-multzoak deskribatzen ditu. Ingelesezko (VCTK eta LibriTTS (Zen *et al.*, 2019)), portugesezko (TTS-Portuguese (Casanova *et al.*, 2022a)) eta frantsesezko M-AILAB (Solak, 2017) datu-multzoak berrerabili ziren. Euskarari dagokionez, bi datu multzo sartu ziren: OpenSLR-76(**?**) eta TTS-DBEU (Sainz *et al.*, 2012). Espainierarako, *TTS-DBES* eta ELRA-TC (Sainz *et al.*, 2012) erabili ziren, *TTS-DBES*-ko hiztunak TTS-DBEUn daude. Gainera, galizierarako (OpenSLR-77 (**?**)) eta katalanerako (OpenSLR-69 (**?**)) datu multzoak gehitu ziren. Ikaskuntza-prozesua ez zailtzeko, gaztelaniaren askotariko euskalkiak ez sartzea erabaki zen, eta gaztelaniazko aldaera aukeratu zen. 3 taulan, p-TTS ereduaren datu-multzo bakoitzaren xehetasunak laburbildu dira.

3. Taula: Datu-multzoen xehetasunak

Datu-multzoa	Hizkuntza	Grabazioak	Hizlariak	Emakumeak	Gizonak	Denbora	Tamaina
VCTK	Ingelesa	44,453	110	63	47	6:12:53	3.6 GB
LibriTTS	Ingelesa	155,471	1,620	889	731	249:52:01	29 GB
TTS-Portuguese	Portugesa	3,625	1	0	1	9:38:45	1,1 GB
M-AILAB	Frantsesa	90,321	5	2	3	173:46:04	19 GB
TTS-DBEU	Euskara	20,698	14	7	7	26:57:27	6,4 GB
OpenSLR-76	Euskara	7,136	52	29	23	9:01:56	2,0 GB
TTS-DBES	Gaztelania	16,158	5	3	2	19:05:28	4,5 GB
ELRA-TC	Gaztelania	5,432	1	1	0	9:48:13	2,2 GB
OpenSLR-69	Katalana	4,240	36	20	16	4:55:23	1,1 GB
OpenSLR-77	Galiziera	5,587	44	34	10	6:45:47	1,5 GB

3.4.3 Emaitzen analisia

P-TTS sistemaren errendimendua balioesteko, giza ebaluazioetan oinarritutako metrikak erabili dira. Metrika horiek funtsezko bi alderditan zentratzen dira: hizketaren naturaltasuna eta jatorrizko hiztunaren ahotsaren antzekotasuna. Jarraian, lortutako emaitzak zehazten dira:

3.4.4 Hizketaren naturaltasuna (MOS - Mean Opinion Score)

Sortutako hizketaren naturaltasuna MOSaren (Mean Opinion Score) bidez ebaluatu zen. Bertan, ebaluatzaileek ahotsaren kalitatea 1etik 5erako eskalan kalifikatu zuten. Emaitzek erakusten dutenez, gure ereduak 4 puntu inguru lortu zituen bi hizkuntzetan (euskaraz eta gaztelaniaz), eta horrek naturaltasun-maila handia adierazten du. Puntuazio horiek *ground truth* (grabazio errealak) puntuazioen azpitik dauden arren, hobekuntza esanguratsua dira aurreko esperimentuekin alderatuta.

4. Taula: Itzulpen-ereduen ebaluazio metriken konparaketa.

Sarrera	MOS
Euskara (TTS-DBEU)	3,926 ± 0,157
Gaztelania (TTS-DBES)	3,990 ± 0,136

3.4.5 Jatorrizko hiztunarekiko antzekotasuna (Sim-MOS)

Jatorrizko hiztunaren ahotsaren antza neurtzeko, Sim-MOSa erabili zen, ereduak hiztunaren ezaugarri bereziak zein ondo erreproduzitzen dituen ebaluatzeko. Kasu honetan, gure ereduak 4 puntu baino gehiago lortu zituen bi hizkuntzetan, eta horrek hiztunaren ahotsa imitatzeko gaitasun ona adierazten du (de Zuazo, 2023).

5. Taula: Itzulpen-ereduen ebaluazio metriken konparaketa.

Sarrera	Sim-MOS
Euskara (TTS-DBEU)	4,015 ± 0,190
Gaztelania (TTS-DBES)	4,170 ± 0,150

3.4.6 Ebaluazioa Zero-Shot hiztunekin

Datuen multzoan sartutako hiztunez gain, zero-shot hiztunekin burututako ereduaren errendimendua ebaluatu zen (entrenamendu-eremutik kanpo). Emaitzek erakusten dutenez, ereduak naturaltasun- eta antzekotasun-maila handia du, baita hiztun ezezagunekin ere, baina puntuazioak apur bat jaitsi dira, datu guztien hiztunekin alderatuta (de Zuazo, 2023).

6. Taula: Naturaltasunaren ebaluazio metriken konparaketa (Zero-Shot).

Sarrera	MOS
Euskara	3,850 ± 0,223
Gaztelania	3,964 ± 0,167

7. Taula: Antzekotasunaren ebaluazio metriken konparaketa (Zero-Shot).

Sarrera	Sim-MOS
Euskara	3,302 ± 0,220
Gaztelania	4,009 ± 0,173

3.5 Web erakuslea eta errendimendu-metrikak

Speech-to-Speech (S2S) sistemaren web-interfazea garatu da erabiltzaile-esperientzia intuitiboa eta eraginkorra eskaintzeko, ahots eleanitzak pertsonalizatzeko eta sortzeko. Horretarako, Gradio *framework*-a erabili da, erabiltzaile-interfaze interaktiboak sortzeko eta API RESTful bat integratzeko aukera ematen duena, ereduaren instantzia bat altxatzean. Gradiok (Abid *et al.*, 2019) ilara-kudeaketa dinamikoa barne hartzen du, sartzen diren eskaeren erabilera eraginkorra ziurtatzen duena, bereziki erabilgarria ereduak eskaera bakoitza prozesatzeko denbora esanguratsua behar duenean. Funtzionaltasun horri esker, eskaerak ordenan ilarakatzen eta prozesatzen dira, sistema saturatzea saihestuz. Oinarri horren gainean, web orri pertsonalizatua diseinatu da, HTML, CSS eta JavaScript bezalako teknologia estandarrak erabiliz. Gradioko APIarekin elkarreragiten du, eskaerak bidaltzeko eta emaitzak modu dinamikoan ikusteko.

8. Taula: Audioaren luzeraren araberako erantzun-denboren ebaluazioa.

Audioaren Luzera	Erantzun-denbora
10 segundo	15,04 segundo
30 segundo	24,37 segundo
1 minutu	24,67 segundo

8 taulan ikus daitekeenez, sistema gai da 30 segundo baino gehiagoko iraupena duten audioetarako denbora errealeko erantzuna sortzeko. Hala ere, audio laburragoen kasuan, karga-denborak latentzia bat sartzen du, erantzuna denbora errealean lortzea eragozten duena. Jarraian, S2S interfazearen irudi bat erakusten da.

2. Irudia: S2S Interfazea

4 Ondorioak

HiTZketan proiektuak bere helburuak bete ditu, euskararen eta gaztelaniaren arteko ahotsetik ahotserako itzulpen sistema integratua garatuz. ASR, MT eta p-TTS teknologiak konbinatuz, hizketaren transkripzioaren zehaztasuna hobetu eta ahots pertsonalizatua sortu da, zero-shot learning teknikak erabiliz. Gainera, puntuazio- eta kapitalizazio-modulua inplementatu da eta web erakusle bat sortu da, erabiltzaileek sistema probatu ahal izateko.

5 Etorkizunerako planteatzen den norabidea

Ondoren, HiTZketan proiektuaren etorkizunerako planteatzen diren hobekuntza hauek jasotzen dira.

- Webgunearen backenda optimizatzea, luzera handiagoko audioak prozesatzeko eta latentzia murrizteko, horrela itzulpenen abiadura hobetuz.
- Arkitekturaren fase guztietan eredu aurreratuenak eta eraginkorrenak ezartzea, hobekuntza berriak sortu ahala.
- Web interfazea eguneratzen jarraitzea, intuitiboagoa eta erabiltzaileentzat eskuragarriagoa izan dadin.
- Tresna modu erosoagoan eta eskuragarriagoan erabiltzea ahalbidetuko duen aplikazio mugikor bat garatzea.

Erreferentziak

Abid, Abubakar, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, eta James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. In *Proceedings of the ICML Workshop on Human in the Loop Learning (HILL 2019)*, Long Beach, USA. ICML.

Casanova, Edresson, Arnaldo Candido Junior, Christopher Shulby, Frederico Santos de Oliveira, João Paulo Teixeira, Moacir Antonelli Ponti, eta Sandra Aluísio. 2022a. Tts-portuguese corpus: A corpus for speech synthesis in brazilian portuguese. *Language Resources and Evaluation* 56.1043–1055.

——, Julian Weber, Christopher D. Shulby, Arnaldo Candido Junior, Eren Gölge, eta Moacir A. Ponti. 2022b. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2709–2720. PMLR.

de Zuazo, Xabier, 2023. *Basque and Spanish Multilingual TTS Model for Speech-to-Speech Translation*. University of the Basque Country - UPV/EHU tesia. Advisors: Dr. Ibon Saratxaga, PhD; Dr. Eva Navas, PhD. Licensed under a Creative Commons Attribution-ShareAlike 4.0 International license.

——, Eva Navas, Ibon Saratxaga, eta Inma Hernáez. 2025. Whisper-lm: Improving asr models with language models for low-resource languages. *HiTZ - University of the Basque Country - UPV/EHU* . Technical documentation and source code available at `http://www.github.com/[anonymized]/whisper-lm`.

Hualde, José Ignacio, eta Jon Ortiz de Urbina. 2011. *A Grammar of Basque*, volume 26 of *Mouton Grammar Library*. Berlin, Germany: Walter de Gruyter.

Sainz, Iñaki, Daniel Erro, Eva Navas, Inma Hernáez, Jon Sanchez, Ibon Saratxaga, eta Igor Odriozola. 2012. Versatile speech databases for high quality synthesis for basque. In *Proceedings of LREC*, 3308–3312.

Solak, I., 2017. The m-ailabs speech dataset. `https://www.caito.de/2019/01/the-m-ailabsspeech-dataset`.

Zen, Heiga, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, eta Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint* .

6 Eskerrak eta oharrak

Lan honek EHUk sustatutako HiTZketan (COLAB22/13) proiektuaren eta Espainiako Eraldaketa Digitalerako eta Funtzio Publikoko Ministerioa eta Europar Batasuneko NextGenerationEU Suspertze, Eraldaketa eta Erresilientzia Planeko ILENIA proiektuaren (2022/TL22/00215335) finantzazioa jaso du.

Eskerrak eman nahi dizkiegu proiektuaren ebaluazioan eta hobekuntzan parte hartu duten erabiltzaileei ere.