IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

VI. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

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

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Neurona-sareetan onarritutako sintesi-ahots pertsonalizatuen hobekuntzen garapena eta ebaluazioa

Iñigo Hierro Muga,
Ibon Saratxaga Couceiro,
Inma Hernáez Rioja,
Ander Arriandiaga Laresgoiti
eta *Victor Garcia Romillo*

235-242 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.29

Neurona-sareetan oinarritutako seintesi-ahots pertsonalizatuen hobekuntzen garapena eta ebaluazioa

Iñigo Hierro Muga, Ibon Saratxaga Couceiro, Inma Hernáez Rioja,
Ander Arriandiaga Laresgoiti, Victor Garcia Romillo
HiTZ Zentroa -Aholab, Euskal Herriko Unibertsitatea (UPV/EHU)
inigoh@aholab.ehu.eus

Laburpena

Adimen artifizialaren garapenarekin batera, neurona-sareetan oinarritutako ahots-sintesi modelo berriak garatu dira, aurreko sintesi teknologiekin lortzen diren emaitzak hobetuz. Sintesi-ahotsen pertsonalizazioa sintesi-ahotsa entrenatzeko beharrezkoak diren baliabide kopurua oso altuari aurre egiteko erabiltzen da, sintesi-ahotsak lortzeko prozesu errazagoa eskainiz. Artikulu honetan artearen egoeraren Text-to-Speech (TTS) modelo baten pertsonalizazioen ebaluazioa zein modeloa hobetzeko proposamenak aurkezten dira eta hobekuntzak egiteko jarraitutako metodologia azaltzen da.

Hitz gakoak: Ahotsaren sintesia, Text-to-Speech, neurona-sareak, sintesi-ahotsen pertsonalizazioa.

Abstract

With the growth of Artificial Intelligence neural network based speech synthesis models that improve the results of previous have been created. The personalization of speech synthesis is a solution to the high number of resources needed to train speech synthesis that offers a new approach to getting synthetic voices. In this paper, we present the evaluation and improvements of the personalization of a state of the art Text-to-Speech (TTS) model.

Keywords: Speech synthesis, Text-to-Speech, neural networks, personalization of synthesis voices.

1. Sarrera eta motibazioa

Ahotsa gizakiok daukagun komunikazio-tresnarik garrantzitsuena da. Zerbait hutsala edo sinplea iruditu arren, tresnarik adierazkorrena eta eraginkorrena da. Historian zehar ahotsa imitatzen dituzten makinak sortzeko saiakerak egon dira. XVIII. mendean, ahotsa sorrarazi zezaketen makina mekanikoak agertu ziren eta geroztik, garapen itzela egon da gaur egungo gailu elektronikoek eskaintzen dituzten ahots-sintesi sistemetara iritsi arte. Ahotsaren sintesia edo TTS (*Text-to-Speech*) edozein testu ahots seinale estandarrean bihurtzean datza (Ning et al., 2019).

Ahotsaren sintesirako teknologiak gizakien arteko komunikazioa errazteko edo ahalbidetzeko erabili daitezke. Ezintasunak dituzten pertsonentzat tresna oso baliagarria izan daiteke. Adibidez, hitz egiteko zailtasunak dituzten pertsonen eta zeinu-hizkuntza ez dakiten pertsonen arteko komunikazioa errazten du. Hori gutxi balitz, Alboko Esklerosi Amiotrofikoa duten gaixoei eta laringektomia bat izan duten pertsonei ahotsa erabiliz komunikatzeko aukera eskaintzen die (Raman et al., 2021)(Serrano et al., 2019).

Merkatuan ahotsaren sintesirako sistema ugari existitzen dira, eta sistema eragile eta teknologia enpresa gehienak punta puntako TTS sistema propioak eskaintzen dituzte. TTS sistema horietan eskaintzen diren ahots generikoen kopurua oso murriztua izaten da. Noizbehinkako erabiltzaileentzat ez da ezinbesteko ahots generikoen eskaria zabala izatea, arreta transmititzen den mezuan soilik jartzen delako. Aldiz, TTS-ak eguneroko bizitzan komunikazio tresna bezala erabiltzen dituzten pertsonentzat, hala nola, mintzamena galdu duten pertsonentzat, gustuko duten ahots sintetiko pertsonala izatea garrantzitsua da. Sintesi-ahotsa komunikazio tresna bat izateaz gain, hizlariaren identitatea eraikitzen laguntzen baitu (Lavan et al., 2019).

Ahotsaren sintesirako modelo bat sortzeko grabazio denbora asko behar da, gutxienez 3-4 ordu inguru. Horretaz gain, grabazioak kalitate oneko ekipoekin eta akustika ona duten inguruneetan egin behar dira, sintesi modeloetan ahal den neurrian zarata txikiena sartzeko eta ondorioz, kalitate oneko sintesi-ahotsa lortzeko. Hori dela eta, lehenago aipatutako ahots pertsonalak lortzeko prozesua zaildu egiten da inbertitu behar den denbora eta ekipamenduaren aldetik. Arazo honi aurre egiteko sintesi-ahotsen pertsonalizazioa deritzon teknika erabiltzen da. Teknika hau,

grabazio denbora nahikoarekin sortu egin den ahots sintetikoa lagin urriak dituen beste ahots batera egokitzean datza, ahots sintetikoak sortzeko beharrezkoak diren baliabideak murriztuz.

Honen harira, Euskal Herriko Unibertsitateko Aholab ikerketa-taldeak AhoTTS izeneko sintesi sistema eleaniztuna garatu du, hainbat plataformetan funtzionatzen duena. Honekin batera, AhoMyTTS[1] izeneko ahots-banku eta ahotsak pertsonalizatzeko sistema garatu dute (Hernaez et al., 2021). Alde batetik, sistema hau edozein gaixotasunagatik hitz egiteko gaitasuna galdu duten pertsonentzako ahots-bankua da. Bestetik, ahotsaren pertsonalizazio sistema bezala funtzionatzen du. Hau da, edonork bere ahotsa dohaintzat eman dezake ahots horrekin hizkera galdu duten pertsonentzako sintesi-ahots pertsonalizatuak sortzeko.

Gizartearen beste hainbat arlotan gertatu den bezala, adimen artifiziala bete-betean sartu da sintesi-ahotsen teknologietan eta neurona-sareetan oinarritutako sintesi-arkitekturak garatu dira. Hauek aurreko sintesi teknologien emaitzak hobetzen dituzte eta sintesi naturalagoak ematen dituzte. Ez hori bakarrik, neurona-sareek eskaintzen duten *fine-tuning* teknika sintesi-ahots pertsonalizazioan ere arrakastaz aplikatu da (Iman et al., 2021).

Honekin guztiarekin, artikulu honetan neurona-sareetan oinarritutako VITS (*Variational Inference with adversarial learning for end-to-end Text-to-Speech*) (Kim et al., 2021) TTS modeloaren sintesi-ahots pertsonalizatuen hobekuntzak proposatzen dira.

2. Arloko egoera eta ikerketaren helburuak

2.1 TTS neuronalak

Emaitza oneko lehenengo ahots-sintesi sistemak ordenagailuen sorrerarekin etorri ziren. Sistema hauek formanteen sintesia, artikulazio-sintesia eta kateatze-sintesia ziren. Ondoren, ikasketa automatikoaren garapenarekin, ahotsaren parametroen aurresatean oinarritzen diren sintesi parametriko estatistikoak jaio ziren. Geroztik, neurona-sareen garapenarekin neurona-sareetan oinarritutako sintesi sistemak (TTS neuronalak) sistema nagusiak bihurtu dira (Tan et al., 2021).

Neurona-sare sakonen gorakadak asko bultzatu ditu TTS sistemak, sistema tradizionalekiko ahots sintetikoen kalitatea eta naturaltasuna hobetuz. Normalean neurona-sareetan oinarritutako TTS sistemak bi moduluz osatuta daude: modulu akustiko neuronala, ezaugarri linguistikoetatik (fonemak eta grafemak adibidez) ezaugarri akustikoak sortzen dituen modulua, eta *vocoder* neuronala (audio sortzailea). Modulu akustiko neuronala baino lehenago gehienetan TTS neuronaletan integratuta ez dagoen testu analisirako modulu bat erabiltzen da (testu normalizazioa, fonemen bihurketa…). Gaur egungo ikuspegirik arrakastatsuena kodetzaile-deskodetzaile ikuspegia da (Ren, et al., 2020) (Mehta et al., 2024). Honetan, kodetzaileak sarrerako testua errepresentazio latentean bihurtu eta deskodetzaileak aldagai latenteak espektrogrametan bihurtzen ditu. Ondoren, *Generative Adversarial Network* (Kong et al., 2020) edo *Flow* (Prenger et al., 2019) tekniketan oinarritutako *vocoder* neuronalek espektrogramak audio seinale bihurtzen dituzte.

TTS neuronalen testuinguruan sintesi-ahotsa terminoa ahots baten grabazioekin entrenatua izan den TTS modelo bati egiten dio erreferentzia. Entrenamendu fasean TTS-ak osatzen duten neurona-sareek balio zehatz bat hartzen dute, ondoren, sintesi fasean edozein sarrerako testuarekin grabazioko ahotsaren antza duen ahots artifiziala sortzeko.

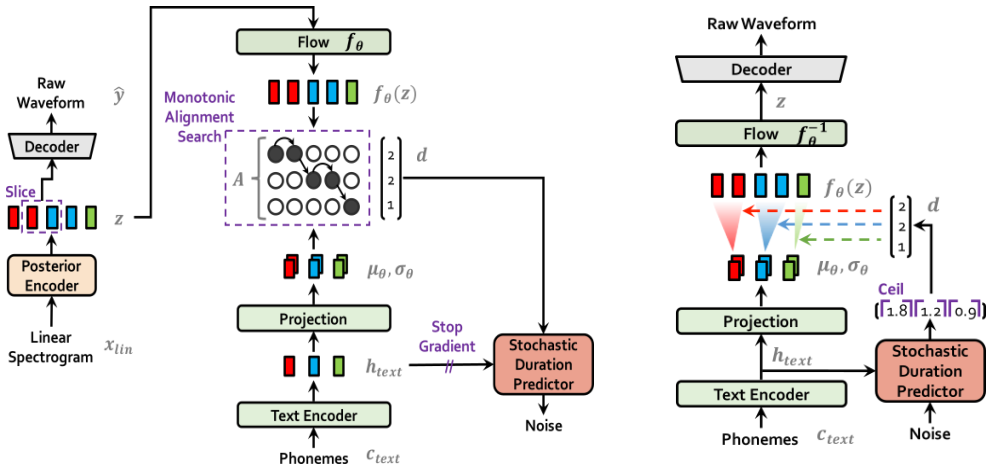
2.2 VITS

Ikerketa honetan erabilitako sintesi sistema VITS izan da, muturretik muturrerako TTS sistema paraleloa. Bi moduluko sistemak entrenatzeko entrenamendu sekuentziala behar da. Bigarren modulua lehengo moduluaren irteerarekin entrenatzen da, entrenamendu prozesua zailduz eta problemak egoteko probabilitatea igoz. Sistema honek modelo akustikoa eta *vocoder*-a bateratu

[1] AhoMyTTS web orrialdea: https://aholab.ehu.eus/ahomytts/

ditu muturretik muturrerako (*end-to-end*) arkitektura bat proposatuz. Ez hori bakarrik, sintesi denbora murrizteko helburuarekin modelo ez-autoregresibo bat proposatzen da. Horrela, laginketa paraleloa ahalbidetzen da gaur egungo prozesadore paraleloen gaitasun osoa erabiltzeko aukera emanez (Kim et al., 2021). 1. irudian VITS sistemaren eskema erakusten da entrenamendu- (ezkerreko atala) zein sintesi-fasean (eskumakoa).

1. irudia. VITS eskema orokorra. (Kim et al., 2021).



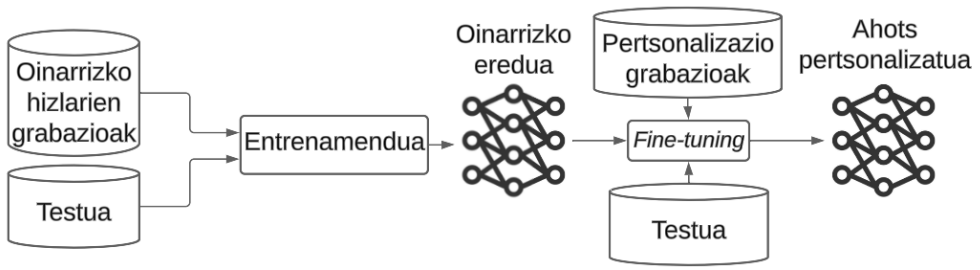
Modeloaren arkitektura azkenengo denboraldian guztion ahotan dagoen adimen artifizial sortzailean oinarritzen da, hain zuzen ere, *Variational Autoencoder*-etan (Kingma eta Welling, 2013). *Variational Autoencoder-a* modulu akustikoa eta *vocoder*-a lotzeko erabiltzen da. Horri, *normalizing flows* eta *adversarial training* teknikak gehituta, adierazpen ahalmen handiko eta kalitate altuko ahots-seinaleak lortzen dira. Testuinguruaren eta emozioaren arabera ahotsaren erritmoa eta abiadura duten aldagarritasuna ikasteko VITS modeloak *stochastic duration predictor* modulu berria inplementatzen du.

2.3 Ahotsen sintetikoen entrenamendua eta pertsonalizazioa

Neurona sare bat entrenatzeko bi era nagusi daude, sarea hutsetik entrenatzea edo, jadanik beste entrenamendu datu desberdinekin entrenatua izan den sare bat hartzea eta sarearen entrenamenduarekin jarraitzea entrenamendu datu berriekin (sarearen egokitzapena, *fine-tuning*). Zerotik egiten den entrenamenduan (*training from scratch*), neurona-sarearen pisuei ausazko balio bat ematen zaie eta datu sorta batekin optimizatzen dira. Aldiz, *fine-tuning* egitean, lehenago egindako entrenamendu batetik lortutako neurona-sareen parametro optimizatuak berrerabiltzen dira (sareko pisuak, adibidez) eta datu sorta berri batekin entrenamendua jarraitzen da. Horrela, beste entrenamendu batean ikasitakoa guk nahi dugun zeregin berrira egokitzen dugu.

TTS neuronaletan ere neurona-sareak egokitzeko teknikak aplikatu daitezke. Artikuluaren sarreran azaldu den moduan, TTS modelo baten entrenamendurako datu sorta on bat lortzea zaila izaten da. Horregatik, *fine-tuning* teknikak eskaintzen dituen abantailetaz baliatuz, ahots pertsonalak entrenamendurako datu sorta murriztuarekin sortu daitezke. TTS modeloaren egokitzapen hauei ahotsaren pertsonalizazioak deituko diegu hemendik aurrera. Pertsonalizazio prozesuan oinarrizko ereduaren neurona-sareak ahots berrira egokitzen dira, ahotsaren ezaugarri akustiko guztiak inplizituki pertsonalizatuz. 2. irudian ahots pertsonalizazioaren prozesua erakusten da.

2. irudia. Ahots pertsonalizazioaren prozesua



Oinarrizko eredua kalitate oneko eta grabazioa ordu nahikorekin entrenatu egin den ahotsari deritzogu. Oinarrizko eredua pertsonalizaziorako ahotsaren datu sortarekin egokitzen denean ahots pertsonalizatua deritzoguna lortzen da.

2.4 Ikerketaren helburuak.

Lanaren helburu nagusia neurona-sareetan oinarritutako ahots-sintesirako modeloen hobekuntzak egitea da, foku nagusia pertsonalizazioetan jarriz. Horretarako, helburu nagusia bi azpi-helburuetan banatu da. Alde batetik, VITS sistemak punta-puntako sintesia egin arren, esaldi batzuen aurrean emaitza arraroak ematen ditu. Hori dela eta, VITS TTS modeloaren entrenamenduak eta ebaluazioak egin dira, TTS-ak dituen erroreak konpontzeko eta pertsonalizazioan erabiliko diren kalitatezko oinarrizko eredu sendoak sortzeko. Bestalde, pertsonalizazioak hobetzeko xedearekin, oinarrizko eredu desberdinek pertsonalizatutako ahotsetan daukaten eragina aztertu da, pertsonalizatu nahi den ahotsaren arabera egokiena den oinarrizko eredua hautatzeko.

3. Ikerketaren muina

Ikerketa bi ataletan banatu da, proposatutako bi azpi-helburuak erdiesteko. Hasteko, sintesi ahotsen ebaluazioa eta akatsak konpontzeko hobekuntza batzuk probatu dira. Ondoren, proposatutako hobekuntza berriekin oinarrizko ereduak eta pertsonalizazioak entrenatu dira eta oinarrizko ereduen eragina aztertzeko lehentasun test bat egin da.

3.1 Oinarrizko ahots sintetikoen akatsen analisia

Hasierako ebaluazioa

Hasierako ebaluazio egin aurretik oinarrizko eredu berriak entrenatu dira 1. taulan erakusten den corpus-a erabilita. Corpus hau Aholab ikerketa taldearen eskuetan dauden grabazio materialak dira (Sainz, et al., 2012). Guztira 4 oinarrizko eredu entrenatu dira: gizonezkoa, G1 ahotsarekin, emakumezkoa, E1 ahotsarekin, E1 eta G1 ahotsekin genero-aniztun eredua eta gizonezko ahots guztiekin eredu konbinatua.

1. taula. Erabili den Corpus-a

Kodea	Corpusaren tamaina	Esaldi kopurua	Kodea	Corpusaren tamaina	Esaldi kopurua
G1	05:01:33	3798	E1	04:44:29	3798
G2	00:37:47	500	E2	00:39:51	500
G3	00:46:08	500	E3	00:41:34	500
G4	00:37:47	500	E4	00:38:45	500
G5	00:35:57	500			

Hasierako ebaluazioa egiteko bakarrik “Gizona” eta “Emakumea” ereduak ebaluatu dira, beste bi ereduek (“Genero-aniztuna” eta “Gizonezko konbinatua”), hizlari desberdinen konbinaketa bat izanik, ez dute hizketa ulergarria ateratzeko gaitasuna eta ezin dira sintesian erabili. Ebaluazioa egiteko 20 esaldiko sorta aukeratu da. Aukeratutako esaldietatik asko hurrengo ezaugarriak dituzte: Esaldi laburrak, hitz propioak, ‘rr’ - ‘j’ - ‘tr’ fonemak eta fonemak esplosiboak. Aurreko ezaugarriak dituzten esaldiek TTS arkitektura askotan erroreak izaten dituztelako aukeratu dira. Ebaluazioatik hurrengo ondorioak atera dira: hitz bakarren esaldien inferentziak ulergaitzak dira, harridura markarekin soinu ulergaitzak sortzen dira eta esaldi amaieran naturalak ez diren mozketak agertzen dira. Ez hori bakarrik, errore hauek oinarrizko ereduetan zein pertsonalizazioetan agertu dira.

Datuen ugaritzea

Hitz bakarreko esaldien eta harridura esaldien inferentzietan agertzen diren erroreak ikusita, oinarrizko ereduak entrenatzeko erabili den corpusa aztertu egin da. Grabaziorik luzeena 10,9

segundokoa da eta laburrena 2,4 segundokoa. Horretaz aparte, corpus osoan ez da aurkitu harridura-markarik. Hortaz, oinarrizko ereduen corpusetan esaldi laburren eta harridura marken gabezia dagoela ondorioztatu da eta horrek, pertsonalizatutako ahotsen sintesian harridura markek eta esaldi laburrek arazoak ematea eragiten du. Gabezia honi aurre egiteko corpusak esaldi laburrekin eta harridura markekin aberastu egin dira. Entrenamenduko datuen tamaina handitzeko bi irtenbide daude. Alde batetik, hasierako entrenamenduko grabazioak sortzeko erabili d iren hizlariekin esaldi sorta berria grabatu daiteke. Aukera honek nahiz eta emaitza hoberenak eman, kostu handiena duena da. Beste aldetik, datuen ugaritze (*data augmentation*) teknikak erabili daitezke. Adimen artifizialaren munduan oso erabilia den prozesura da eta hainbat eratara egin daiteke.

Kasu honetan, "TTS-by-TTS: TTS-driven Data Augmentation for Fast and High-Quality Speech Synthesis" (Hwang et al., 2021) lanean proposatzen den prozedura jarraitu da gure entrenamenduko datuen tamaina handiagotzeko. Aholab taldeko beste proiektu batean, ahots berberekin FastSpeech2 (Ren, et al., 2020) TTS modeloak entrenatu dira eta horiek ez dituzte hitz bakarreko esaldiekin eta harridura markekin VITS aurkezten dituen erroreak ematen. Hori dela eta, FastSpeech2-ren sintesi-ahots horiek erabili dira entrenamenduko datuen tamaina handiagotzeko. Audio berriak sortzeko izenordainak, erakusleak, galdegaiak eta hitz egiterakoan esamolde edo esaera bezala erabiltzen diren izenak, adberbioak eta aditzak hartu dira, guztira 452 hitzeko zerrenda sortuz. Horretaz gain, harridura marka duten 99 esaldi gehitu dira.

Oinarrizko ereduak entrenamendurako datu sorta berriarekin entrenatu eta gero, esaldi laburrak ulergarriak dira eta harridura marketan ez dira soinu arraroak agertzen.

Isiluneen mozketetako babes-tarteen doikuntza

Esaldien amaieran gertatzen diren ebaketa ez normalak konpontzeko xedearekin, entrenamendurako audioen prozesamenduari erreparatu diogu. Audioek entrenamendura sartzeko prest egoteko hainbat prozesamendu teknika jasaten dituzte. Horietako bat hasierako eta amaierako isiluneen ebaketa da. Isiluneak audioa grabatzerako orduan agertzen dira eta funtsezkoa da horiek kentzea isiluneak oso luzeak izan daitezkeelako eta ereduak hori ikas dadila ekidin nahi delako. Isilune hauek modu automatizatuan kentzeko Silero-VAD izeneko VAD (*Voice Activity Detector*) modelo bat erabiltzen da. VAD-aren lana esaldia audio grabazioaren zein unetan hasten den esatea da. Emaitzak bisualki aztertu ondoren, VAD modeloaren zehaztasuna perfektua ez dela ikusi da eta hortaz, ebaketa prozesuan ahotsaren seinalearen laginak ebakitzen direla aurkitu da. Ondorioz, TTS modeloaren entrenamendura ebakita dauden esaldiak sartu dira inferentzian sintetizatutako esaldietan amaierako mozketak eraginez.

Arazo hau konpontzeko prozesamenduan hasierako eta amaierako babes-tarteak uztea adostu egin da. Horrela, VAD-aren zehaztasuna perfektua ez izan arren, ahots seinalea entrenamenduan osorik sartzen dela ziurtatu da. Babes-tartearen luzera ideala aurkitzeko "Emakumea" oinarrizko ereduko hiru eredu entrenatu dira 0,05 segundoko, 0,1 segundoko eta 0,3 segundoko babes-tartearekin. Hiru ereduak entrenatu ondoren, 0,05 segundoko babes-tartea uztea adostu da.

3.2 Oinarrizko ereduen eragina ahots pertsonalizazioan

Oinarrizko ereduen entrenamendua

Ahotsaren pertsonalizazio prozesu honela ulertu daiteke: oinarrizko ereduaren entrenamenduan ahotsaren ikasketa nagusia gertatzen da, ondoren, pertsonalizazioan ahotsari azkenengo zertzeladak ematen zaizkio. Hori dela eta, pertsonalizazioak egiteko oinarrizko ereduaren aukeraketa aproposa egitea funtsezkoa da. Pertsonalizatu nahi den ahotsari hobekien egokitzen zaion oinarrizko eredua aldatu daiteke ahotsaren ezaugarrien arabera.

Hori aztertzeko 4 oinarrizko eredu berri entrenatu dira aurreko ataletako hobekuntzak inplementatuz. Entrenatutako oinarrizko ereduen entrenamendu datu sorta 1. taulan erakusten da. Aipatu behar da, datuen ugaritzea atalean deskribatutako prozeduraren bitartez, hizlari bakoitzaren corpusa handitu dela. Ezaugarri desberdineko oinarrizko ereduak entrenatu dira, ahalik eta ahots-eremu handiena ukitzeko (2. taula).

2. taula. Oinarrizko eredu berriak

Eredua	Data
Gizona	G1
Emakumea	E1
Genero-aniztuna	G1 + E1
Konbinatua	G1+G2+G3+G4+G5+E1+E2+E3+E4

Entrenamendura sartzen diren audioak 16 bit-eko bit-emaria daukate 22050 Hz-ko laginketa maiztasunarekin. Hobekuntza atalean proposatu bezala 0,05 segundoko babes-tartea utzi da. Horretaz gain, audioei Short-Time Fourier transformatua (STFT) aplikatu zaie 1024 lagineko Fast Fourier transformatua (FFT), 256 lagineko saltoak eta 1024 lagineko leiho tamaina erabiliz.

Entrenamenduan erabilitako batch tamaina 32-koa izan da eta *bucketing* teknika erabili da. Teknika honekin tamaina antzeko audioak batch berdinetan jartzen dira audio luzerak berdintzeko erabiltzen den betegarria murriztuz eta hortaz, entrenamenduan erabiltzen den memoria optimizatuz. Ereduak 400.000 eta 500.000 arteko iterazioz entrenatu dira. Entrenamendua gelditzeko erabilitako irizpidea guztiz subjektiboa izan da, iterazio desberdinen emaitzak entzun eta kalitatea hobetzen ez dela nabaritu denean entrenamendua gelditu da. Ikasketa automatikoko beste aplikazio batzuetan ez bezala, TTS sistemen entrenamenduan zaila da sistemaren errendimendua metriken bitartez aztertzea, gizakion entzumen-sentikortasuna ez datorrelako bat modeloak eman ahal dituen errore metrika objektiboekin (adibidez, Mean Squared Error (MSE)).

Pertsonalizazioak

Behin oinarrizko ereduak sortu direla pertsonalizaziorako ahotsak hautatu dira. Gizonezko 3 ahots, emakumezko 3 ahots, 2 ume ahots eta emoziodun ahots bakarra. Ahots bakoitzak 100 esaldiko corpusa dauka, guztira 10 minutuko grabazioekin. Ahots horiek AhoMyTTS ahots-bankutik atera dira.

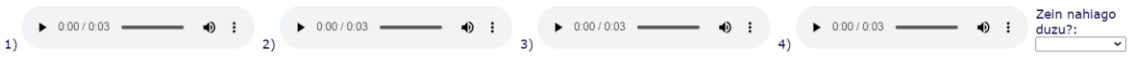
Entrenamenduko datuak oinarrizko ereduek izandako aurre-tratamendu berdina jaso dute. Pertsonalizazioaren kasuan *bucketing* teknika alde batera uztea adostu da esaldiak antzeko luzerak dituztelako. Horretaz gain, hainbat entrenamendu egin ostean heuristikoki adostu da 4000 iterazio ondoren pertsonalizazioak gelditzea, 2 ordu inguru.

Lehentasun testa

Oinarrizko eredu desberdinek pertsonalizatutako ahotsetan daukaten eragina aztertzeko ebaluazio-test bat zabaldu da. Audio ebaluazioak egiteko erabiliena izaten den metodoa *Mean Opinion Score* (MOS) teknika da, zeinetan audioen kalitate subjektiboaren batezbesteko emaitza kuantifikatua lortzen den. Hala ere, gure helburua 4 ereduetatik hobekien egokitzen den oinarrizko eredua aurkitzea izanik, lehentasun testa bat egitea adostu da entrenatutako ereduen artean hobeto moldatzen den eredua aurkitzeko. Test mota hauetan eskainitako aukeretatik gustukoena aukeratzen da.

Lehentasun testa web bidez zabaldu da hurrengo formatuarekin: ebaluatzaileak 4 audioen artean gustukoena aukeratzen du (3. irudia). Audio laukote bakoitzak esaldi berdina dauka eta audio bakoitza sintesi-ahots desberdinak erabiliz sintetizatu da. Laukote bakoitzaren 4 sintesi-ahotsak pertsonalizaziorako ahots berdinarekin entrenatu dira, baina 4 oinarrizko eredu desberdinetatik moldatuz.

3. irudia. Lehentasun-testaren adibidea



Lehentasun testean 30 esaldi erabili dira eta 36 sintesi-ahots ebaluatu dira (4 oinarrizko ereduko pertsonalizaziorako 9 ahots), guztira 1080 audio sortuz. Test-a luzeegia ez izateko, test bakoitzean bakarrik 32 audio laukote ebaluatu dira, pertsonalizatutako ahots bakoitzetik 4 audio laukote.

Emaitzen analisia

Testean 22 partaide izan ditugu, guztira 796 audio laukoteen ebaluazioa lortuz. 3. taulan testaren emaitzak ikusi daitezke. Gehien aukeratu diren audioak eredu konbinatutik pertsonalizatutako ahots pertsonalizatuenak izan dira. 4.irudian emaitzak era grafikoan adierazten dira. Lortutako emaitzek hizlari desberdinen ahotsen konbinazioekin sortutako oinarrizko ereduak ahots eremu zabalago batera hobeto egokitzen direnak direla erakusten dute. Pertsonalizatutako ahotsak banaka aztertuz gero, 7 ahotsetan eredu konbinatua nagusiena izan da, gainerako 2 ahotsetan emakumezko eredua izan da nagusiena (4.taula).

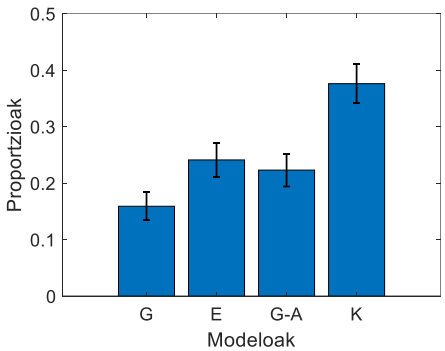
3. taula. Emaitza orokorrak

	Gizona	Emakumea	Genero-aniztuna	Konbinatua	Guztira
N	126	191	177	298	792
Proportzioa	0,16	0,24	0,22	0,38	1

4. taula. Ahotsaren araberako emaitzak

AHOTSA	Gizona	Emakumea	Genero-aniztuna	Konbinatua
Gizona1	28	5	23	**32**
Gizona1	21	14	17	**36**
Gizona3	20	7	18	**43**
Emakumea1	9	28	17	**34**
Emakumea2	8	**31**	21	28
Emakumea3	5	27	21	**35**
Umea1	22	19	20	**27**
Umea2	4	24	25	**35**
Emozioduna	9	**38**	15	28

4. irudia. Emaitzak era grafikoan



4. Ondorioak

Artikulu hau urrats garrantzitsua da neurona-sareetan oinarritutako sintesi-ahots pertsonalizazioaren garapenerako. Proposatutako hobekuntzak AhoTTS sisteman integratu diren VITS TTS ahots sendoak entrenatzeko erabili dira. Horretaz gain, lehentasun testarekin, orokorrean hobekien egokitzen den oinarrizko eredua aurkitu da. Eredu hori eredu konbinatua da

(gizonezko eta emakumezko hainbat hizlarirekin entrenatua) eta hemendik aurrera, AhoMyTTS pertsonalizazio sisteman erabiliko da ahots pertsonalizatuen oinarrizko eredu gisa.

5. Etorkizunerako planteatzen den norabidea

Etorkizunera begira, lan honetatik ateratako ondorioak eta konpontze teknikak gaztelaniazko ahots sintetikoetan aplikatuko dira. Hau da, datuen ugaritze eta babes tarteen tratamendua. Horrez gain, atea irekitzen da beste TTS neuronalak pertsonalizatzera eta berriro oinarrizko ereduen azterketa bat egitera, VITS TTS-en atera diren ondorio berdinak ateratzen diren ala ez ikusteko.

6. Erreferentziak

Hernáez Rioja, I., Navas Cordón, E., Saratxaga Couceiro, I., & Sánchez de la Fuente, J. (2021). Ahots sintetiko pertsonalizatuak: esperientzia baten deskribapena. *EKAIA EHUko Zientzia eta Teknologia aldizkaria, ale berezia 2021*, 173–194. https://doi.org/10.1387/ekaia.22077

Hwang, M.-J., Yamamoto, R., Song, E., & Kim, J.-M. (2021). TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis. *In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6598-6602). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414408

Iman, M., Arabnia, H. R., & Rasheed, K. (2023). A Review of Deep Transfer Learning and Recent Advancements. *Technologies*, *11*(2), 40. https://doi.org/10.3390/technologies11020040

Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, IEEE.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*. https://doi.org/10.48550/arXiv.1312.6114

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis. *Advances in Neural Information Processing Systems, 33*, 17022–17033. https://doi.org/10.5555/3495724.3497152

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review, 26*(1), 90–102. https://doi.org/10.3758/s13423-018-1497-7

Mehta, S., Tu, R., Beskow, J., Székely, É., & Henter, G. E. (2024). Matcha-TTS: A fast TTS architecture with conditional flow matching. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 11341–11345.

Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L.-J. (2019). A review of deep learning based speech synthesis. *Applied Sciences*, *9*(19), 4050. https://doi.org/10.3390/app9194050

Prenger, R., Valle, R., & 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)*, 3617–3621. https://doi.org/10.1109/ICASSP.2019.8683143

Raman, S., Sarasola, X., Navas, E., & Hernaez, I. (2021). Enrichment of oesophageal speech: Voice conversion with duration-matched synthetic speech as target. *Applied Sciences, 11*(13), 5940. https://doi.org/10.3390/app11135940

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). FastSpeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*. https://doi.org/10.48550/arXiv.2006.04558

Sainz, I., Erro, D., Navas, E., Hernáez, I., Sánchez, J., Saratxaga, I., & Odriozola, I. (2012). Versatile speech databases for high quality synthesis for Basque. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3308–3312. http://www.lrec-conf.org/proceedings/lrec2012/pdf/126_Paper.pdf

Serrano, L., Raman, S., Tavarez, D., Navas, E., & Hernaez, I. (2019). Parallel vs. non-parallel voice conversion for esophageal speech. *Proceedings of Interspeech 2019*, 4549–4553. https://doi.org/10.21437/Interspeech.2019-2194

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*. https://doi.org/10.48550/arXiv.2106.15561

7. Eskerrak eta oharrak

Lan honetarako finantzazioa honako iturrietatik lortu da: Espainiako Eraldaketa Digitalerako eta Funtzio Publikoko Ministerioa eta Europar Batasuneko NextGenerationEU Suspertze, Eraldaketa eta Erresilientzia Plana (ILENIA project, 2022/TL22/00215335).