EKAIA (2024), artikulua prentsan/article in press.
https://doi.org/10.1387/ekaia.26342
Behin-behineko bertsioa (euskara-orrazketaren faltan).

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN: 0214-9001
e-ISSN: 2444-3255

Hizketa-interfaze isilak: ahotsaren sorrera bioseinaleetatik. Seinale elektromiografikoen bidezko azterketa

(Izenburua Ingelesez) Silent Speech Interfaces: generation of speech from biosignals. A study with electromyographic signals.

Eder del Blanco, Inge Salomons, Eva Navas, Inma Hernáez, Ibon Saratxaga, Jon Sánchez*

HiTZ Basque Center for Language Technology, University of the Basque Country, Spain

LABURPENA: *Hizketa-interfaze isilak (SSI) bioseinale ez-akustikoetatik abiatuta mintzamena deskodetu dezaketen gailuak dira, eta mintzamen-urritasunak dituzten pertsonen komunikazioan dituzte aplikazioak. Lan honetan, EMG-seinaleetan oinarritutako SSI bat diseinatzeko jarraitutako metodologia, garatutako ereduaren arkitektura eta entrenamendua, eta emaitzen kalitatea ebaluatzeko metodologia deskribatzen dira. Grabazio saioen arteko aldakortasuna murrizteko, hau da, sentsoreen posizioa bezalako faktoreek eragindako EMG-seinalearen aldaketak murrizteko, 3D maskarak erabili izan dira, kokapen finkoa ziurtatzeko. Esaldien sintesia ebaluatu da, bai eredua entrenatzeko erabili diren saioetako adibideak erabiliz, bai ikusi gabeko saioetako adibideak erabiliz, ereduaren errendimendua egoera bietan konparatzeko. Emaitzek erakusten dute garatutako sistemak saio-aldaketarekiko sendotasuna duela, hainbat saiotan EMG-seinaleak lortzeko jarraitutako estrategiari esker. Lan hau aitzindaria da gaztelaniarako SSI bat garatzen eta saioen arteko mendekotasunaren azterketa zehatza egiten, aurreko azterlanetan alderdi hauek landu ez baitira.*

HITZ GAKOAK: hizketa-interfaze isilak, sare neuronal sakonak, transformers, EMG, seinale elektromiogramikoak, ahots-bihurketa, EMG-ahots bihurketa, laringektomizatutako pertsonak×

ABSTRACT: *Silent Speech Interfaces (SSIs) are devices capable of decoding speech from non-acoustic biosignals, offering a means of communication for individuals with speech disabilities. This paper presents the methodology for developing an SSI based on electromyography (EMG), detailing the model architecture, training process, and evaluation approach. To address variability between sessions, caused by factors such as sensor positioning, 3D masks were employed to ensure consistent sensor placement. Sentence synthesis was evaluated using examples both from the training sessions and from unseen sessions, allowing comparison of the model's performance in both scenarios. Results indicate that the system demonstrates robustness to session variability, attributed to the acquisition strategy used for EMG signals across different sessions. This work is innovative in its development of an SSI for Spanish and its extensive analysis of session dependency, aspects that have not been previously explored in similar studies.*

KEYWORDS: silent Speech Interfaces, Deep Neural Netwoks, transformers, EMG, Electromyographic Signals, Voice Conversion, EMG-to-Speech conversion, Laryngectomees.

***Harremanetan jartzeko/ Corresponding author**: Eder del Blanco, Komunikazioen Ingeniaritza, Bilboko Ingeniaritza Eskola, Torres Quevedo Ingeniaria Enparantzia, 48013 Bilbao. https://orcid.org/0000-0001-6510-778X, eder.delblanco@ehu.eus

Nola aipatu / How to cite: del Blanco, Eder; Salomons, Inge; Navas, Eva; Hernáez, Inma, Saratxaga, Ibon, Sánchez, Jon (2024). << Hizketa-interfaze isilak: ahotsaren sorrera bioseinaleetatik. Seinale elektromiografikoen bidezko azterketa >>, Ekaia, xx, xx-xx. (https://doi.org/10.1387/ekaia.26342)

Jasoa: maiatzak 22, 2024; Onartua: abaneduak 2, 2024

ISSN 0214-9001-eISSN 2444-3225 / © 2024 UPV/EHU

1. SARRERA

Hizketa-interfaze isilak (*Silent Speech Interfaces*, SSI) [1], [2], [3] etorkizun oparoko aukera gisa agertu dira ahozko komunikazioa berreskuratzeko, hizketa deskodetuz hizketa ekoizteko prozesuan sortzen diren bioseinale ez-akustikoetatik abiatuz. Bioseinale mota anitz ikertu dira SSIrako: ahots-traktuaren irudiak [4], artikulografia elektromagnetikoa (artikuladore-mugimenduen atzemate magnetikoa) [5], elektromiografia (EMG) (aurpegiko muskuluen jarduera atzematea da, gainazaleko elektrodoak erabiliz) [6], artikuladoreen mugimenduak detektatzen dituzten ultrasoinuak [7], eta elektroentzefalografia (EEG) [8] edo magnetoentzefalografia (MEG) [9] burmuineko elektrizitate-aktibitatea harrapatzeko. SSIek seinale akustikoren beharrik gabe hizketa sortzeko aukera ematen dutenez, irtenbide berri bat eskaintzen dute mintzamen-desgaitasuna duten pertsonen komunikazio-gaitasunak berrezartzeko.

Bi ikuspuntu desberdin proposatu dira hizketa deskodetzeko bioseinaleak abiapuntua izanik: bioseinaleak testu bihurtzea eta zuzeneko ahots-sintesia. Lehenengo ikuspuntuan hizketa-ezagutze automatikorako algoritmoak erabiltzen dira *(Automatic Speech Recognition*, ASR) bioseinale-datuekin entrenatuak, sarrera-datu horietatik dagokion testua deskodetzeko. Ondoren, behar izatekotan, testu-ahots bihurgailu bat (*Text-to-Speech*, TTS) erabil daiteke ahotsa sintetizatzeko deskodetutako testuarekin. Bigarren hurbilketan sistema bat entrenatuko da bioseinaleetatik zuzenean ahotsa sortzeko gai dena, testuaren deskodifikazioa gauzatu barik.

Interfaze isilen aplikazioen artean ondorengoak daude, besteak beste: ingurune zaratatsuetan ahozko komunikazioa ahalbidetzea, kanpoko pertsonek entzun behar ez dituzten elkarrizketetan pribatutasuna ematea eta ahozko ezgaitasuna duten pertsonei komunikatzea bermatzea, modu eraginkor eta independenteagoan. Azken aplikazio hori da artikulu honetan aurkeztutako bi ikerketa proiektuen (amaitu den ReSSInt proiektuaren eta abian dagoen DeepRestore izenekoaren) lanaren motibazioa. Proiektu horietan, laringektomizatutako pertsonei ahozko komunikazioa berrezartzeko, EMG- eta bideo- seinaleetatik abiatuta ahotsa sortzea eta ezagutzea ikertzen da.

Lan honetan, orain arte lortutako emaitza garrantzitsuenak aurkeztuko ditugu, SSIen egungo egoera deskribatuz (2. atala), sortutako baliabideak (3. atala), ikerketa-ildo hori garatzeko erabilitako metodologia (4. atala), EMG-seinaleak hizketa bihurtzeko esparruan egindako esperimentuak (5. atala) eta emaitzetatik lortutako ondorioak azalduz. Azken atalak proiektuaren etorkizuneko lan ildoak planteatzen ditu.

2. EGUNGO EGOERA

Atal honetan, elektromiografia aurkezten dugu, seinaleak eskuratzeko zenbait metodo eta EMG-seinaletan oinarritutako SSIekin lotutako beste ikerketa batzuek aurpegiko zein muskulu hartu duten kontuan deskribatzen dugu. Ondoren, EMG-seinaletatik abiatutako zuzeneko hizketa-sorreraren egungo egoera esploratzen da. Helburu hau zaila da, EMG-seinaleak deskodetzea eta kalitatezko hizketaren sintesia oso konplexuak direlako, baina denbora errealean funtzionatzen duten sistemak lortzeko aukera ematen du.

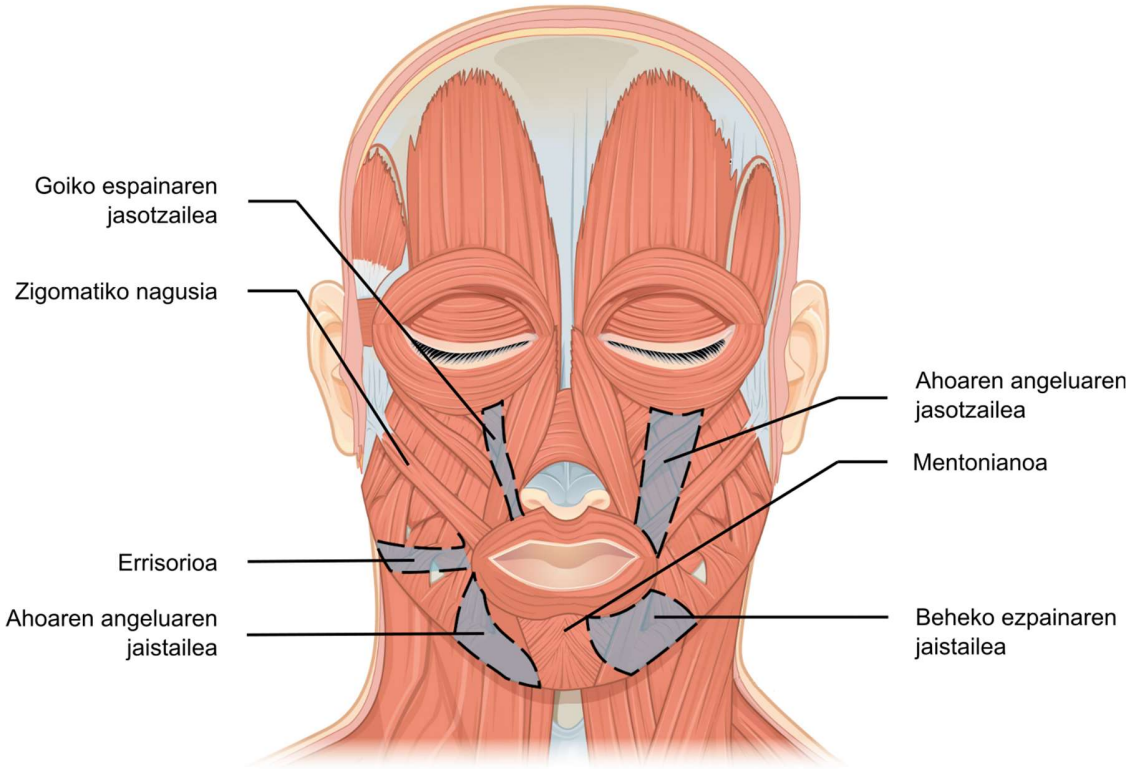
2.1 EMG-seinaleen eskuratzea

Elektromiografia muskulu-jarduera atzemateko erabiltzen da, izenak berak adierazten duenez: muskulutako (mio) elektrizitatearen (elektro) erregistroa (grafia). Muskuluak uzkurtu egiten dira burmuineko seinaleek aktibatutako bulkada elektrikoen ondorioz, eta bulkada horiek falta direnean erlaxatzen dira. Bulkada elektriko horiek, EMG-seinaleak, gailu baten bidez erregistra daitezke, muskuluan txertatutako (EMG inbaditzailea; iEMG) edo larruazalari itsatsitako (azaleko -*surface*-EMG; sEMG) elektrodoen bidez. EMG-seinaleak adierazten du muskuluan elektrizitatea dagoen ala ez. Gure lanetan sEMG aukeratu dugu, bere izaera ez-inbaditzailea dela eta.

EMG-seinaleak lortzeko sentsoreak bi konfigurazio desberdinetan koka daitezke: monopolarrean edo diferentzialean. Monopolarrean, erreferentziazko elektrodo bat behar da, muskulu-jarduerarekin lotutako bulkadarik espero ez den lekuan kokatzen dena, belarriaren lobuluan adibidez. Gero, erreferentziazko elektrodoaren seinalea kendu egiten zaio aztertu nahi dugun muskuluan dagoen elektrodo monopolarrak hartutako seinaleari. Diferentzialean, berriz, bi puntutan neurtutako seinaleen arteko aldea neurtzen da, elektrodo bipolarrak erabiliz. Bi neurketa-puntuek kanal bat osatzen dute, dela elektrodo bipolarren bi poloen artean, dela erreferentziazko elektrodoaren eta elektrodo monopolarraren artean, dela seriean jarritako bi elektrodoen artean.

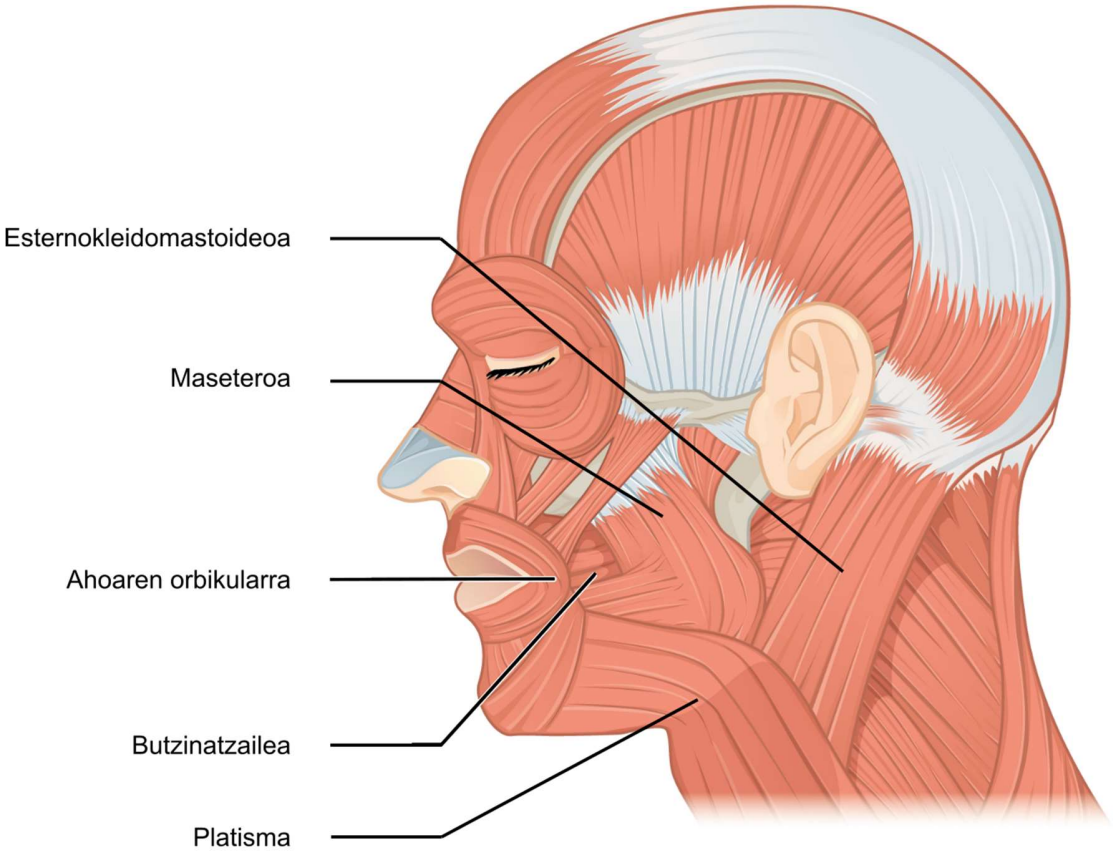
Hainbat lanetan ikertu da nola lortu daitezken aurpegiko muskuluen EMG-seinaleak, SSItarako aplikaturik, azterketaren helburuaren araberako muskulu-hautaketa bat erabiliz. Azterketa askok muskulu espezifikoetan jartzen dute arreta, eta gehienek bost muskulutako multzo bat aukeratzen dute: ahoaren angeluaren jasotzailea (*levator anguli oris*), zigomatiko nagusia (*zygomaticus major*), platisma (*platysma*), ahoaren angeluaren jaistailea edo depresorea (*depressor anguli oris*) eta muskulu digastrikoaren aurreko sabela (*anterior belly of the digastric*), mihiarekin erlazio handiena duen azaleko muskulua, alegia [10], [11], [12], [13]. Beste azterketa batzuek muskulu desberdinak erabiltzen dituzte,

hala nola, butzinatzailea (*buccinator*), ahoaren orbikularra (*orbicularis oris*), mentonianoa (mentalis), goiko ezpainaren jasotzailea (*levator labii superioris*), milohioideoa (*mylohyoid*), esternokleidomastoideoa (*sternocleidomastoid*) edo irri-muskulua edo errisorioa (*risorius*) [14], [15], [16], [17], [18], [19]. 1. irudian muskulu horien kokapena ikus daiteke. Beste lan batzuek ez dira muskulu espezifikoetan ardazten, baizik eta erregio anatomiko mugatuetan [20], [21], [22], [23]. Horretarako, dentsitate handiko sentsoreak[1] edo matrizeak erabil daitezke [24]. Beste lanetan, erregio anatomiko zabal bat ikertzen dute, banakako sentsoreak [25] edo elektrodo-matrizeak erabiliz, gero kanal garrantzitsuenak hautatzeko [26].

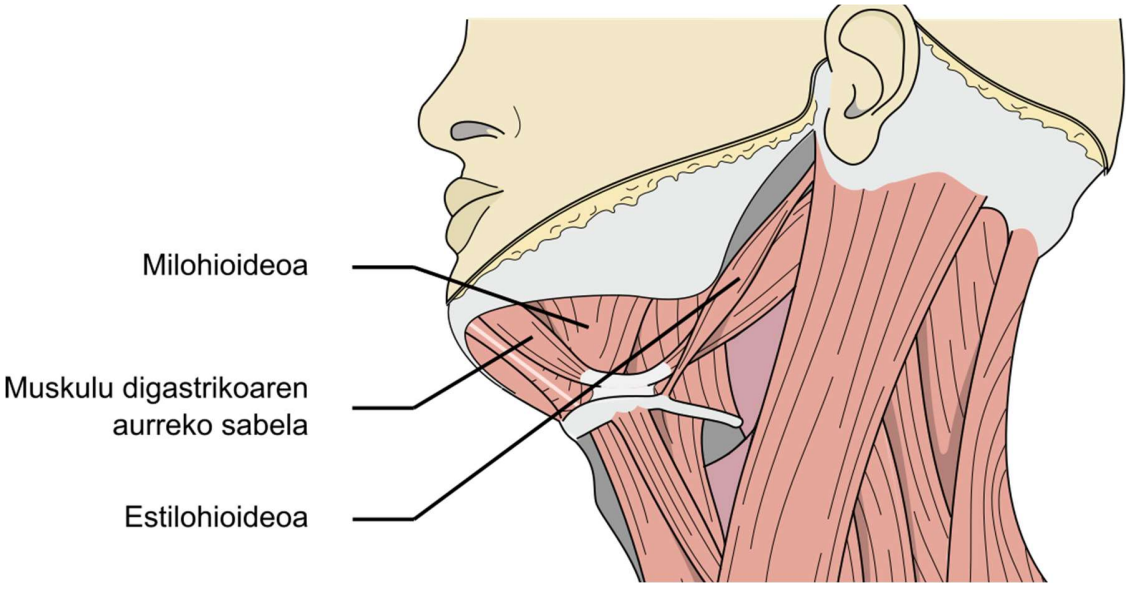


1.a irudia. Aurpegiko muskuluak, aurrez aurreko bista. Urdinez markatutako muskuluak irudian ageri ez diren azpiko muskuluak dira.

[1] Dentsitate handiko elektrodoak matrizearen konfigurazio batean hainbat elektrodo dituzten sentsoreak dira, muskulu-jardueraren bereizmen espazial handia lortzeko.



1.b irudia. Aurpegiko muskuluak, alboko bista.



1.d irudia. Aurpegiko muskuluak, alboko bista zeiharra.

2.2 *EMG-to-Speech*

EMG-seinaleen bidezko zuzeneko ahots-sorrera ez dago hain aztertua, atazaren zailtasuna dela eta. Sistema hauen ebaluazio-metrika gisa, sortutako ahots-seinaleak ASR sistema batetatik pasatzean lortutako transkripzioen errore-neurriak erabili dira, hala nola hitzen errore-tasa (*Word Error Rate*, WER) edo karaktereen errore-tasa (*Character Error Rate*, CER), eta baita sortutako seinalearen eta ahots-seinale objektibo baten arteko espektro-distantziaren neurketak, hala nola mel espektrogramen arteko distantzia edo distortsio cepstrala [27]. Distortsio cepstralak audio seinale baten cepstruma[2] eta erreferentziarenaren arteko aldea neurtzen du. Lehenengo saiakeretako batean, EMG-seinaletik abiatuta espektro-inguratzailea lortzeko ahots-bihurketa aplikatu zen, eta, f0 eta energia-balio naturalekin konbinatuta, % 15,7ko WERa lortu zen [28]. Ondoren, sare neuronalak, gaussiar-nahasteen ereduak eta unitateak hautatzeko teknikak aplikatu ziren hizketa entzungarria gauzatzen den bitartean sortutako EMG-seinalea ahots bihurtzeko, EMG seinaleak sentsore-matrize batez eskuratuz eta EMG-seinaleekin batera grabatutako audioa erabiliz. Hala, sare neuronalen kasuan, 5,21 dB-ko distortsio cepstrala zuten seinaleak lortu ziren, baina sortutako seinale gehienak ez ziren ulergarriak [29]. [23] erreferentzian, hizketa isilaren bidez sortutako EMG seinaleak audio entzungarrian bilakatzeko *transformer* sare-arkitektura erabiltzea proposatzen da. Metodo horrekin sortutako ahotsak, esatari baten 18,6 grabazio ordurekin entrenatuta, ASR sistema baten bidez ebaluatuz % 68ko WERa lortzen du, giza transkripzioen kasuan % 75era igotzen dena. Ondoren, eredu hori are gehiago hobetu zen CNN bat erabiliz EMG-seinaleen ezaugarri hoberenak automatikoki ateratzeko; hala, % 44,8ko WERa lortu zen ASR sistema batekin eta % 32,3koa giza ulergarritasunerako [23]. [30] erreferentzian sekuentziatik sekuentziarako[3] eredua aurkeztu zen, hizketa isileko EMG-seinaleen eta ahozko hizketaren arteko lerrokaduratik iraupenari buruzko informazioa ateratzen duena, 6 txinera hiztunen datu kopuru txiki batekin entrenatuta (ordubete inguru). Eredu horrek % 22ko batezbesteko CERa lortu zuen ASRrekin, eta % 6,4koa giza transkripziorako. Beste lan batek ahots-bihurketa teknika aplikatzea planteatu zuen esatari desberdinei dagokien ahotsa sortzeko, hizketa isileko EMG-seinale batetik abiatuta. Seinale hori ahots-unitateetan kodetzen da, eta xede-esatariaren errepresentazio (*speaker embedding*) batekin batera pasatzen da deskodetzaile batera, bertan parametro akustikoak sortzeko. Azkenik, horiek ahots bihurtzen dira *vocoder* baten bidez [31]. Lan horretan, batezbesteko % 42,22ko WERa lortu da. Nahiz eta EMG-seinaletatik ahotserako bihurketa zuzenaren emaitzak hobetu egin diren urtez urte, oraindik erronka

[2] Cepstrum-a Fourier-en alderantzizko transformatua espektroan aplikatzearen emaitza da.
[3] Sekuentziatik sekuentziarako eredu bat, *seq2seq* ere deitua, sarrerako datuen sekuentzia bat irteerako datuen sekuentzia bihurtzen duen eredu bat da.

ugari daude ebazteko, hala nola emaitzak grabazio-saioarekiko eta entrenatutako pertsonarekiko mendekotasunik ez dituzten sistemak garatzea.

Hemen deskribatzen den lana, funtsean, [23] eta [32] erreferentzietan deskribatutako lanean oinarritzen da. Lan honek oso emaitza onak ditu, eta kodea eta datu-basea ikerketa-komunitatearentzat eskuragarri daude; beraz, abiapuntu egokitzat jo zen gaztelaniarako eskuratutako datuekin esperimentatzeko.

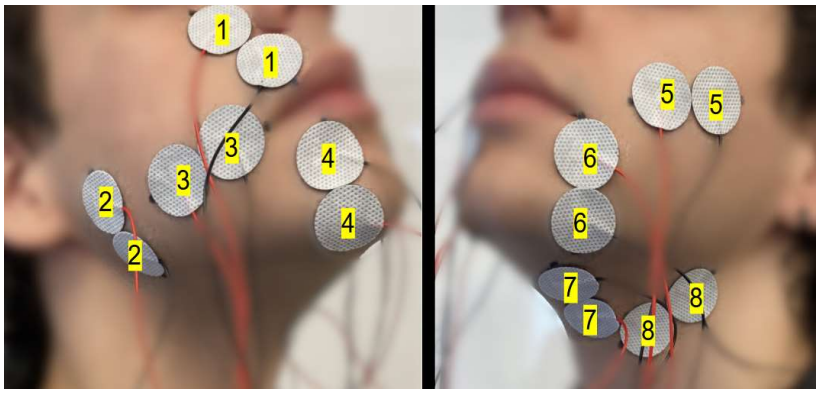
3 ReSSInt-EMG DATU-BASEA

Lan honetarako ReSSInt-EMG datu-basea erabili da, ReSSInt[4] proiektuan garatutakoa. Grabaketak gauzatzeko elektrodo bipolarrak erabili ziren, muskulu bakoitzaren gainean longitudinalki kokatuak, muskulu bakarreko seinalearen erregistroa egiteko, erregio anatomiko bateko muskuluen mugimendu bateratuaren informazioa hartu beharrean. Hizketaren ekoizpen-prozesuan EMG-seinaleak eskuratzeko muskulu egokienei buruzko adostasunik ez dagoenez, analisi pilotua egin genuen, aurpegiko eta lepoko azaleko muskulu garrantzitsu guztiak aztertuz, zereginerako muskulu egokienak identifikatzeko. Azterketa horren ondorioz, 2. irudian ageri den azken konfigurazioa definitu zen, helburu-muskulu hauek dituena (muskulu bakoitzerako kanal bat erabiliz):

- Goiko ezpainaren jasotzailea (1. kanala) (*levator anguli oris*)
- Maseteroa (2. kanala) (*masseter*)
- Errisorioa (3. kanala) (*risorius*)
- Beheko ezpainaren jaistailea (4. kanala) (*depressor labii inferioris*)
- Zigomatiko nagusia (5. kanala) (*zygomaticus major*)
- Ahoaren angeluaren jaistailea (6. kanala) (*depressor anguli oris*)
- Muskulu digastrikoaren aurreko sabela (7. kanala) (*anterior belly of the digastric*)
- Estilohioideoa (8. kanala) (*stylohyoid*)

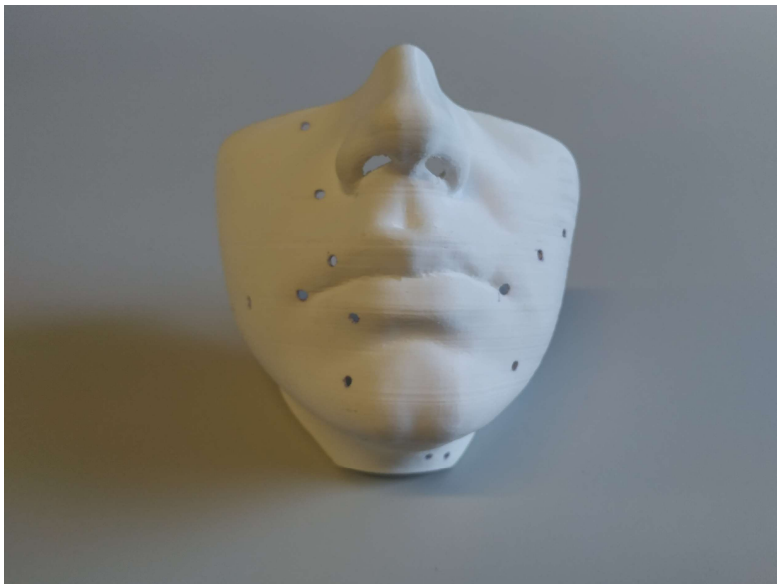
1. irudian erakusten dira muskulu horiek.

[4] http://ressint.eus. 2021/09/23an onartu zuen proiektua UPV/EHUko CEISH erakundeak, 142/2021 aktaz.



2**. irudia**. Grabazioetan erabilitako elektrodoen konfigurazioa, zortzi elektrodo bipolar pareak (zortzi kanal) erakusten dituena. Pare bakoitzak muskulu desberdin bat du xede.

Grabazio-saio bakoitza hasi aurretik, teknikari aditu batek EMG-sentsoreak leku egokietan jartzen zituen esatariaren aurpegia eta lepoa garbitu ondoren. Muskulu bakoitzean lortutako seinaleen ezaugarriak aldatu litezke saio batetik bestera, sentsoreak muskuluaren leku desberdinetan jarrita baleude. Arrazoi horregatik, sensoreak saio guztietan leku berberetan jartzen direla bermatzeak berebiziko garrantzia du, ahal den zehaztasun handienarekin. Horretarako, 3D maskara pertsonalizatua egin zen parte-hartzaile bakoitzarentzat (ikus 3. irudia). Horiek sortzeko, muskuluak lokalizatu eta erreferentzia-puntuak markatu ondoren, 3D eskaner bat erabili zen esatarien aurpegia eskaneatzeko. Maskaran, sentsoreen kokapena markatzeko erreferentzia-puntuak zulatu ziren. Ondorengo saioetan, markak pertsonaren aurpegian margotzen dira zuloak erabiliz, eta elektrodoak horren arabera jartzen dira; hala, posizioak leku berean mantentzen dira saioen artean.



3. irudia. 3D maskara pertsonalizatua. Zuloak erreferentzia-markak aurpegiaren azalean margotzeko erabiltzen dira elektrodoek subjektuaren aurpegian dituzten posizioak aurkitzeko.

Grabazioak grabazio-kabina erdiprofesional eta intsonorizatuan egin ziren. Grabazio-saio guztietan, teknikari aditu batek saio osoa gainbegiratzen zuen, lortutako seinaleak etengabe monitorizatuz, elektrodoak askatzeak, interferentziek, ebakera txarrek edo beste edozein arazok eragindako akatsak saihesteko.

Datu-baseak baditu bai grabazio entzungarriak (*a*), non esatariek aurkezten zaien testua ozenki irakurtzen duten, bai grabazio isilak (*s*), non esatariek pantailan agertzen diren hitzak eta esaldiak artikulatzen dituzten, baina ahotsik sortu gabe. Grabazio-saio batzuek seinale entzungarriak (*a*) besterik ez dituzte, beste batzuek seinale isilak (*s*) baino ez dituzte, eta beste batzuek testu bereko ebakera entzungarriak eta artikulatuak dituzte (*a+s*, testu bera behin ozen eta behin isilik, edo *a+2s*, behin ozen eta birritan isilik grabatzen denean). Azken hauei *saio mistoak* deituko diegu. 1. taulan saio guztien zehaztasunak ageri dira.

1. taula Saio bakoitzean grabatutako corpusa. *a* kodeak adierazten du corpusa modu entzungarrian grabatzen dela, *s* isilean, *a+s* bai modu entzungarrian, bai isilean grabatzen dela, eta *a+2s*, behin modu entzungarrian eta bi aldiz isilean grabatzen dela. Grisez markatutako parteak ez dira erabili deskribatutako esperimentuetan.

Corpus	Saio ID														
	001	002	003	004	005	006	007	008	009	010	011	012	013	014	015
110 BKB	a	a	a	a	a+s	s			s						
100 hitz	a	a	a	a	a+s	s	a+s	a+s	s						
Sharvard 1-100	a	a	a	a	a+s	s	a+s	a+s	s						
Sharvard 101-400	a					s									
Sharvard 401-700		a							s						
AhoSyn 1-150			a												
AhoSyn 151-300				a											
AhoSyn 301-400							a								
AhoSyn 401-500								a							
AhoSyn 501-505										a+2s	a+2s	a+2s	a+2s	a+2s	a+2s
AhoSyn 506-570										a+2s					
AhoSyn 571-635											a+2s				
AhoSyn 636-700												a+2s			
AhoSyn 701-765													a+2s		
AhoSyn 766-830														a+2s	
AhoSyn 896-960															a+2s

1. taulan ikus daitekeenez, grabaketak 15 saioetan egituratu dira (001etik 015era zenbakituta), grabaketen edukiaren arabera: bokal-kontsonante-bokal (BKB) konbinazioak, hitz isolatuak eta corpus desberdinetatik ateratako esaldiak). Ez dugu BKB konbinazioei eta hitz isolatuei buruzko azalpenik

hemen gehituko, deskribatuko diren esperimentuetan erabili ez baitira. Esaldiak Sharvard [33] eta AhoSyn [34] gaztelaniazko corpusetatik hartu dira. Sharvard corpusa 700 esaldiz osatuta dago, eta fonemikoki orekatuta dago. Honek esan nahi du fonemak corpusean agertzen diren maiztasuna gaztelanian naturalki agertzen diren maiztasunaren antzekoa dela. AhoSyn corpusa sortzean, bestalde, gaztelaniaren difono gehienak barnean hartzen saiatu dira. Azken corpus honetan hitz eta izen atzerritarrak gaztelaniazko izenekin aldatu dira, esatariek era berdinean ahoskatuko dituztela ziurtatzeko.

Datu-baseak esatari bat baino gehiago duen arren, datu gehien grabatuta duen esatariaren datuak baino ez dira erabili lan honetan. Esataria 29 urteko gizonezko bat da, 15 saio grabatu dituena. Esaldiak bakarrik kontuan izanda, 4 ordu eta 6 minutu baino gehiago EMG-seinale eta 2 ordu eta 25 minutu ahots-seinale dira, ahots-seinalea grabazio entzungarrietatik soilik lortzen dela kontuan hartuta. 2. taulan saio bakoitzaren iraupen zehatza erakusten da.

2. taula Esaldien grabazioen iraupena saio bakoitzeko.

Saioa	Audio-seinaleen iraupena (mm:ss)	EMG-seinaleen iraupena (mm:ss)
001	16:48	16:48
002	17:29	17:29
003	16:57	16:57
004	19:13	19:13
005	05:19	10:49
007	15:10	20:55
008	14:48	20:51
010	07:02	22:53
011	07:03	22:52
012	06:23	20:03
013	06:30	20:31
014	06:03	18:59

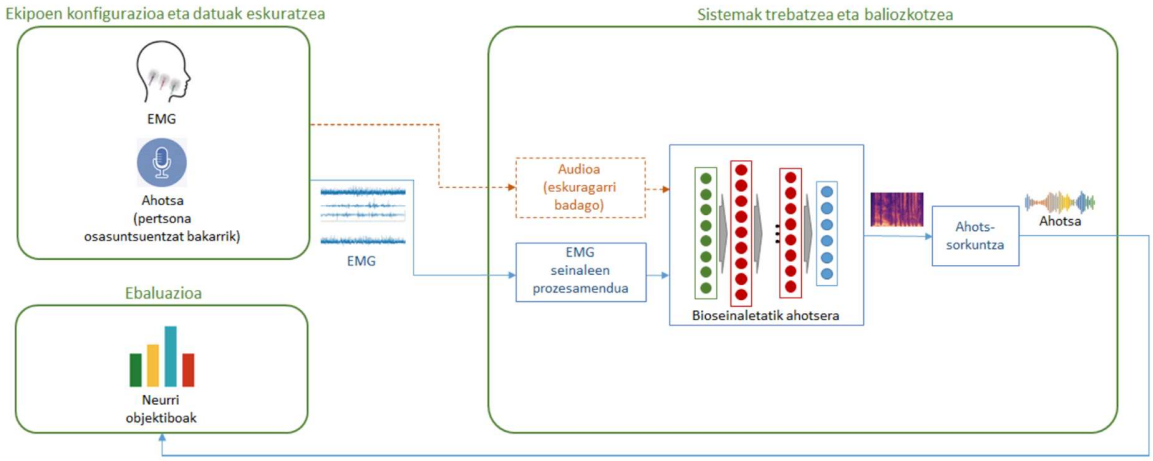
015	05:51	18:25
Guztira	**144:36**	**246:45**

EMG-seinaletatik ahotsa sortzeaz gain, ataza osagarri gisa, fonemen sailkapena egingo da, 4. atalean azalduko dena. Horretarako, beharrezkoa da audio-grabazioen segmentazio bat sortzea, une bakoitzean zein fonema ahoskatzen den jakiteko. Etiketa horiek sailkapena ebaluatzeko erreferentzia gisa erabiliko dira. Hori egiteko, Montreal Forced Aligner (MFA) [35] softwarea erabili dugu, gaztelaniaren 29 fonemako multzoa gehi isiltasun etiketa ezarriz. MFAk audioak berak erabiltzen ditu testuen transkripzio fonetikoekin batera, fonema bakoitza esaldiaren zein denbora-tartetan ahoskatzen ari den estimatzeko.

4 METODOLOGIA ETA SISTEMAREN ARKITEKTURA

4.1 SSI garatzeko jarraitutako metodologia

Hizketa isileko interfazeak garatzeko proposatutako metodologia funtsezko hainbat etapaz osatuta dago, 4. irudian erakusten den bezala. Hasiera batean, arreta fase esperimentalean jarri zen. Fase horretan, besteak beste, seinale elektromiografikoak eta ahotsa grabatzeko ekipoak konfiguratu ziren, parte-hartzaileak lortu eta datuak grabatu ziren. Sistema honen garapenean zehar, eta datuen bilketarekin batera, fonemen sailkapen esperimentuak egin ziren, informazioa behar bezala biltzen ari zela eta EMG-seinaleek gure asmoetarako informazio erabilgarria zeukatela ziurtatzeko [36], [37]. Aurreko atalean deskribatutako datuak eskuratu ondoren, neurona-sareetan oinarritutako algoritmoak erabili dira hizketa isilean lortutako seinale biometrikoetatik zuzenean mintzamena sortzeko. Algoritmo horien entrenamendu-fasean, EMG- eta ahots-seinaleak erabiltzen dira bi seinaleei buruzko informazioa korrelatzeko gai diren ereduak garatzeko. Inferentzia-fasean, bakarrik EMG-seinaleak erabiltzen dira ahots-irteera sortzeko. Algoritmo horien eraginkortasuna zorrotz ebaluatzen da metrika objektiboak erabiliz (fonema-zehaztasuna eta mel distantzia gisakoak). Proba subjektiboak alde batera utzi ditugu, gaur egun ereduaren garapena dagoen atariko fasea dela eta.



4. irudia. EMG-seinaleetan oinarritutako SSIak garatzeko metodologia. Lerro etenak sistemen entrenamendu-faseari bakarrik dagozkio.

4.2 Sistemaren deskribapen orokorra

EMG-seinaleak ahots bihurtzeko erabilitako arkitektura aurretiko lanetan oinarrituta dago [23], [32], [38]. Lan horien gainean doikuntza txikiak egin dira zenbait parametro gure seinaleak eskuratzeko sistemaren ezaugarrietara egokitzeko. Lan honetan, sistema ulertzeko garrantzitsutzat jotzen ditugun alderdiak soilik deskribatuko ditugu. Hurrengo azpiataletan dimensio eta konfigurazio parametroen zehetasunak ematen dira.

5. irudiak erabilitako oinarrizko eskema erakusten du. Ikus daitekeenez, sareak ez du iragartzen zuzenean audio-seinale bat uhin-forma gisa, baizik eta mel espektrograma bat. Ondoren, etapa independente batean, iragarritako mel espektrograma audio-seinale bihurtuko da *vocoder* neuronal baten bidez [39].

Sarea entrenatzeko, sarrera gisa 8 EMG-kanalen seinaleak erabiltzen dira, iragazita eta birlaginduta. Iragazpen prozesuak osagai zuzena, mugimendu geldoak eta interferentzia elektrikoa ezabatzen ditu. Beharrezkoa da EMG-seinalea birlagintzea, gerora audio-seinaleekin ondo lerrokatzeko. 4.2 azpiatalean prozesu hauei buruzko xehetasunak azaltzen dira.

Sarrerako laginak prestatu ondoren, sare konboluzionalen bloke batek erauziko ditu bihurketa-sarean sartuko diren parametroak. Bihurketa-sarea *transformer* arkitektura batean oinarrituta dago [40]. *Transformer*raren azken geruzaren irteera proiekzio linealeko azken etapa batetik igarotzen da, irteerako mel espektrogramaren bektoreari dagokion dimentsiora murrizteko (80 balio trama bakoitzeko).

Arkitekturak fonemen sailkapena egin dezake baita. Horretarako azken etaparen ordez proiekzio linealeko beste geruza bat dago, hizkuntza bakoitzari dagozkion fonema-klaseen artean sailkatzeko probabilitateak sortzen dituena: 30 gaztelaniarako eta 48 ingeleserako. 5. irudian, fonema kopurua *F* aldagaiarekin adierazten da.

Azkenik sistema entrenatzen da, ereduaren emaitzak bi eginkizunetarako optimizatuz: alde batetik, iragarritako mel espektrogramen eta erreferentziazko mel espektrogramen arteko desberdintasuna balioztatzen da, eta beste aldetik fonema-sailkapenean lortutako zehaztasuna.

4.3 Datuen prestaketa

Lehenago aipatu bezala, sarearen entrenamenduan sarrera gisa 8 EMG-kanalen seinaleak erabiltzen dira, iragazita eta birlaginduta. 5. irudiaren nomenklatura jarraituz, *n* esaldi bakoitzaren EMG-seinalearen iraupena da, laginetan adierazita. Noski, esaldi baten 8 kanaletako seinaleek iraupen berbera izango dute. Iragazketak 2 Hz-etik beherako osagaiak ezabatzen ditu 3 ordenako Butterworth goi-paseko iragazki batekin, osagai zuzena eta aldaketa geldoak ezabatzeko. 50 Hz-eko osagai harmonikoak ere iragazten dira, Q = 30-eko Notch iragazki batekin. Bi noranzkoetan iragazten da seinalea, iragazkia oinarrizko maiztasunari eta bere harmoniko guztiei aplikatuz, korronte elektrikoak sar dezakeen zarata garbitzeko.

Ondoren, seinaleak 689,66 Hz-ean birlagintzen dira; izan ere, aurrerago azalduko den bezala, ereduak ezaugarri konboluzionaleko erauzgailu bat du, 8 EMG-seinaleko lagin multzo bakoitzeko trama bat sortzen duena. Birlaginketak bermatzen du 8 laginen iraupena mel espektrogramaren trama-luzerarekin lerrokatuko dela, eta hori 11,6 ms-ko tramekin sortuko da (256 lagin 22.050 Hz-ean), EMG-seinaleen eta audio-tramen arteko korrespondentzia egokia ahalbidetuz. EMG-seinaleak birlagindu ondoren, luzera berria *n'* laginekoa izango da. 5. irudian, *n'* EMG-laginekin lortzen den trama-kopurua *N* aldagaiarekin adierazten da. Irteerako tramaren luzera mel-espektrograma audio bihurtzeko erabiliko den *vocoder*raren beharrizanen arabera ezartzen da.

Hurrengo urratsa ereduaren sarrera gisa erabiliko diren EMG-seinaleak kateatzea da. Entrenamendua esaldi sortatan egiten denez, EMG-seinaleak sekuentzia bakar batean kateatzen dira, 8 kanalak mantenduz. Sekuentzia hori kanal bakoitzeko 1600 lagineko luzera finkoko zatietan banatzen da (ikus 6. irudia). Azken zatiak luzera hori betetzen ez badu, zeroekin osatzen da (*padding*). 5. irudian, urrats honetan lortzen den sekuentzia kopurua *M* aldagaiarekin adierazten da, eta bere balioa sorta bakoitzeko esaldi guztien luzeren araberakoa da. Estrategia horrek esaldi bakoitzean sarrerako sekuentziaren luzera finkoa lortzeko *padding*a sartzeko beharra ekiditen du. Gainera, seinaleak zatika banatzeak

entrenamendua hobetzen du, ereduari zenbait esaldiren testuinguru osoa kentzen baitio, erregularizatzaile gisa jokatuz [32]. Ereduaren diseinuak irteerako trama bakoitzari sarrerako EMG-seinalearen 8 lagin esleitzen dizkionez, irteerak sekuentzia bakoitzeko 200 tramatako luzera finkoa izango du, eta trama horiek berregituratu egingo dira esaldi bakoitzaren jatorrizko luzera duten sekuentziak izateko.

4.4 Ezaugarrien erauzketa eta transformazioa

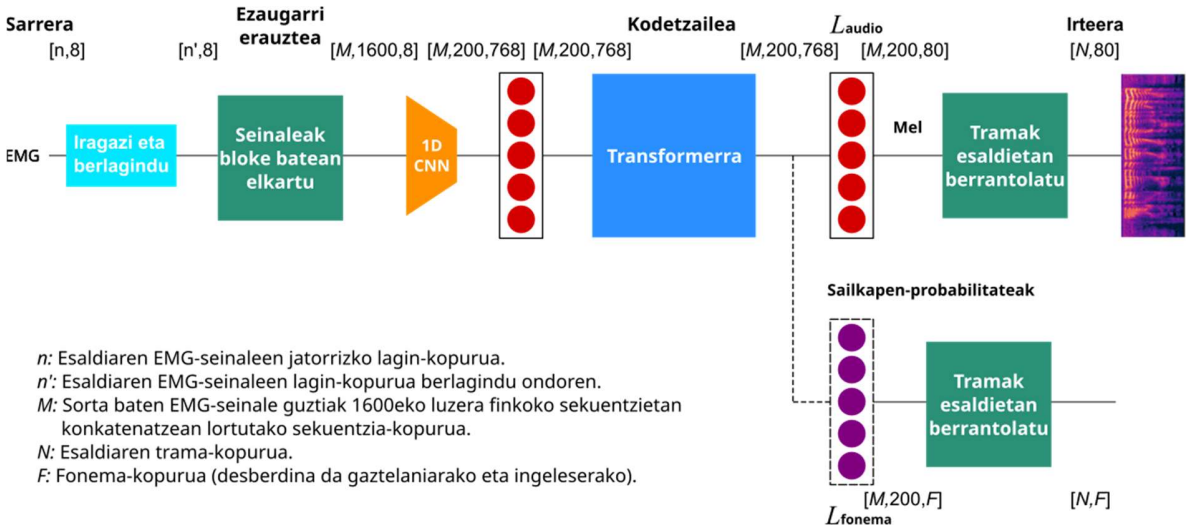
Sarrerako laginak luzera finkoko sekuentziatan prestatu ondoren, sare konboluzionalen multzo batek aterako ditu bihurketa-sarean sartzeko parametroak. Sare konbolutzionalen multzo honek 768 ezaugarri[5] erauziko ditu 8 EMG lagin multzo bakoitzerako, hau da, trama bakoitzerako. Ondoren, erauzitako ezaugarriak proiekzio linealeko geruza batetik pasatzen dira, eta gero bihurketa-sare baten bidez prozesatzen dira. Bihurketa-sarea *transformer* arkitektura batean oinarrituta dago [40]. Bi sareen konposizio- eta konfigurazio-xehetasunak zehazki deskribatzen dira [32] erreferentzian.

Azkenik, *transformer*rak bihurtutako ezaugarriak sartzen dira, alde batetik, mel espektrogramak tramaz-trama iragartzen dituen 80 dimentsioko proiekzio linealeko geruza batera eta, beste alde batetik, trama bakoitzarako 30 klaseen arteko klasifikazio probabilitateak iragazten dituen beste proiekzio linealeko geruza batera.

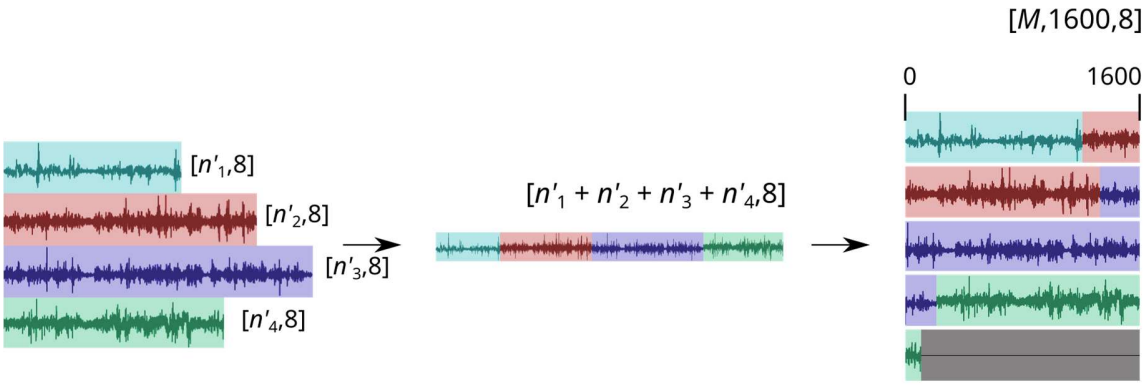
Ereduak luzera finkoko sekuentzia bakoitzerako 200 trametarako iragarpenak egin ondoren, bai mel espektrogramaren tramak, bai trama bakoitzarako egindako klasifikazioak berrantolatzen dira jatorrizko esaldien luzera dituzten sekuentzietan.

Azkenik, behin eredua entrenatuz gero, HiFi-GAN (*High Fidelity Generative Adversarial Network*) [39] *vocoder* bat erabiltzen da, sareak iragarritako mel espektrogrametatik audio-seinale bat sortzeko egokitua. Abiapuntutzat erabili den kodearekin [38] ematen den HiFi-GAN eredua erabili dugu, ingelesez entrenatuta.

[5] *Transformer*raren tamaina aurretiko lanetatik hartzen da.



5. irudia. EMG-seinaleak ahots bihurtzeko sistemaren oinarrizko arkitektura. Irteeran lortutako mel espektrograma ahots-seinale bihurtzen da *vocoder* neuronal baten bidez.



6. irudia. Sorta baten esaldiak luzera berdineko sekuentziatan antolatzeko prozesua. □′□ adibide bakoitzaren EMG-seinaleen luzera adierazten du, lehenengo birlaginketa gauzatu ondoren. *i* esaldiaren indizea adierazten du (irudiaren adibidiezko sortak soilik 4 esaldi ditu). *M* lortutako sekuentzia-kopurua da (adibide honetan, □ = *5*). Adibide bakoitzak 8 kanal ditu, guztiak luzera berberarekin.

4.5 Galera-funtzioaren kalkulua eta entrenamendua

5. irudian ikusten denez, xede-seinale nagusia (*target*) mel espektrograma izango da. Fonema-sailkapen ataza entrenatzeko, fonema-transkripzioetatik lortutako etiketak xedetzat erabiltzen dira. Mel espektrograma eta denborarekiko lerrokatutako transkripzio-fonetikoak sortzeko ahots-grabaketak beharrezkoak direnez, beti erabiltzen dira grabaketa entzungarriak *target* moduan.

Aurretiko lanetan [23], [32] frogatu den bezala, bai modu entzungarrian bai modu isilean ekoiztutako EMG-seinaleekin entrenatu behar da sistema. Sarreratzat isileko-grabaketen EMG seinalea erabiltzen denean, saio berberako grabaketa entzungarri bat, testu-eduki berbera duena, erabiltzen da xedetzat. Esaldi entzungarri hauek *a+s* eta *a+2s* izendatutako esaldi multzoetako *a* parteak izango dira, alegia. Kasu hauetan, ereduaren sarreran erabiltzen den esaldiaren eta xedearen iraupena desberdina da, eta honek konparaketa galarazten du. Modu isilean lortutako EMG-seinaleekin entrenatzea ahalbidetzeko, EMG-seinaleez bestelako iraupeneko audio-seinaleak erabiltzea ahalbidetzen duen estrategia bat bilatu behar da. Aukeratutako metodoa sareak iragarritako audioaren eta erreferentziazko audioaren arteko denbora-lerrokatze dinamikoa (*dynamic time warping*, DTW) egitean datza. Estrategia horren bidez, posible da entrenamendurako erabiltzea bai entzuteko moduan grabatutako esaldietatik bai isilik grabatutako esaldietatik lortutako EMG-seinaleak.

Sistema entrenatzeko erabiltzen den galera-funtzioa 1. ekuazioa adierazten duen moduan definituko da.

1. ekuazioa: $\square_{\square\square\square\square\square} = \square_{\square\square\square\square\square} + \square\square_{\square\square\square\square\square\square}$,

non $\square_{\square\square\square\square\square}$ ereduak iragarritako mel espektrogramaren distortsioa da, $\square_{\square\square\square\square\square\square}$ fonemen sailkapenak eragindako galera da, eta λ parametro bat da, zeinek $\square_{\square\square\square\square\square}$ galera akustikoen eta $\square_{\square\square\square\square\square\square}$ fonema-identifikazio galeren pisuak $\square$ galera osoen barruan doitzeko balio duena. 0,5eko balioa aukeratu da λ-rako, 0, 0,3, 0,5 eta 0,7 balioak probatu ondoren, audioaren galeraren eta fonemaren galeraren arteko oreka ona lortzen duela egiaztatu dugulako.

Mel espektrogramaren distortsioari dagozkion galerak zuzenean lortzen dira sareak iragarritako mel espektrogramaren ($\widehat{\square}[\square]$) eta *target*aren ($A[i]$) arteko distantziatik, 2. ekuazioan azaltzen den bezala.

2. ekuazioa: $\square_{\square\square\square\square\square} = \sum_{\square}^{\square_I} \quad \sum_{\square=\square\square\square(\square)} \quad ||\square[\square] - \widehat{\square}[\square]||$

non *i* eta *j* indizeek mel espektrograma bakoitzaren tramen denbora-posizioa adierazten baitute eta $\square_I$ target espektrogramaren luzera trametan. Indize ezberdinak erabiltzen dira, sareak iragarritako mel espektrogramak eta *target*ak luzera ezberdina izan lezaketelako $\hat{A}[j]$ iragarpen ahoskera isileko EMG-seinaleetatik lortzen denean (*a+s*-ren *s* zatia) eta *target*a ozen egindako ahoskeraren audioa denean (*a* zatia). DTW(x) funtzioak bi indizeen arteko lerrokaketa adierazten du.

Gainera, informazio osagarri gisa, sareari uneko tramari dagokion fonemari buruzko informazioa ematen zaio, eta hari lotutako galerak 3. ekuazioaren bidez kalkulatzen dira:

3. ekuazioa: $\square_{\square\square\square\square\square\square} = -\sum_{\square}^{\square_I} \quad \sum_{\square=\square\square\square(\square)} \quad \square[\square]^{\square}\square\square\square\ \widehat{\square}[\square]$

non $P[i]$ xede-fonemari (*target*) dagokion *one-hot* bektorea da eta $\widehat{\square}[j]$ fonemaren sailkapenari lotutako probabilitate-bektorea. 2. ekuazioan bezala, arrazoi beragatik, konparatu beharreko bektoreek luzera desberdina izateko posibilitateak azaltzen du bi indize desberdin erabiltzea.

1. ekuazioan deskribatu den galera-funtzioa optimizatuz entrenatzen da sarea. Eredua 100 *epoch*en zehar entrenatzen da, AdamW optimizatzailea erabiliz eta ereduaren balioztatzerako erreserbatutako datu-multzoarekin galera txikiena lortzen duen eredua gordez. Sorta-tamaina 32 esaldikoa da. $2{,}5 \cdot 10^{-4}$ ikaste urrats-tamaina erabiltzen da, urrats-tamainaren beroketa lineala inplementatuz (*warm-up learning rate*). Estrategia hori jarraituz, entrenamendua ikaste urrats-tamaina txiki batekin hasten da, eta tamaina hori pixkanaka handitzen da sorta bakoitzarekin, helmugako tamainara iritsi arte, formula honi jarraituz (4. ekuazioa):

4. ekuazioa: $\frac{\square_{\square\square\square\square h} \cdot \square\square_{\square\square\square\square\square\square}}{\square_{\square\square\square\square - \square\square}}$,

non $\square_{\square\square\square\square h}$ uneraino entrenatzeko erabilitako sorta-kopurua den, $\square\square_{\square\square\square\square\square\square}$ helmugako ikaste urrats-tamaina , eta $\square_{\square\square\square\square - \square\square}$ helmugako urrats-neurrira heltzeko behar den urrats kopurua. Ezarritako helmugako ikaste urrats-neurrira behin helduta, hura erdira murrizten da baliozkotze-galera bost *epoch*etan zehar hobetzen ez den bakoitzean.

5 ESPERIMENTUAK

Lan honetan EMG-seinaletik ahotsa sortzeko hasierako esperimentuak deskribatzen dira, ReSSInt-EMG datu-base sortu berriarekin eginak. Esperimentu hauekin ereduaren portaera aztertzen da, esatari bakar baterako, bi baldintzetan: alde batetik saioarekiko mendekotasuneko baldintzetan (SD, *Session Dependent*, hemendik aurrera), eta beste aldetik saioarekiko mendekotasunik gabeko baldintzetan (SI, *Session Independent*).

Saioen arteko aldakortasuna saio batetik bestera aldatu ahal diren alderdien multzoa da, zeintzuek jasotako EMG-seinaleen ezaugarrietan eta kalitatean eragina dituzten. 3. atalean azaldu den bezala, oso garrantzitsua da saio guztietan sensoreak leku berean jartzea, jasotako EMG-seinaleek ahalik eta ezaugarri antzekoenak izan ditzaten. Horretaz gain, badaude saioen arteko aldakortasunean eragina dituzten beste faktore batzuk, kontrolpean ez daudenak. Adibidez, ingurunearen hezetasunak edo

tenperaturak elektrodoen itsasgarritasuna gutxitu ahal dute. Baita esatariearen gogo-aldarteak ahoskatzeko edota artikulatzeko era orokorra aldatu dezake saio batetik bestera. Beraz, espero izatekoa da saioen arteko aldakortasunak ereduen errendimenduan eragina izatea. Eraginaren larritasuna baloratzeko, saioarekiko mendekotasuna kontutan hartzen duten esperimentuak egiten dira.

Esperimentu hauetan, eredu batzuk entrenatzen dira saio guztiak erabiliz bat izan ezik, aldi bakoitzean saio desberdin bat baztertuz. Eredua SI baldintzetan probatzeko, baztertutako saioetatik ateratako esaldiak erabiltzen dira. Bestalde, eredua SD baldintzetan aztertzeko, eredua entrenatzeko erabili diren saioetatik ateratako esaldiak erabiltzen dira eredua probatzeko, betiere xede honetarako gorde diren esaldiak erabiliz, ereduak entrenamenduan inoiz ikusi ez dituenak. Gainera, probatzen diren esaldien testuak ez dira erabili entrenamendurako (hau da, ereduek ez ditute inoiz ikusi probatzen diren esaldien testuak).

Hurrengo ataletan egindako esperimentuak azaltzen dira: 5.1 azpiatalean emaitzen onura neurtzeko erabiltzen diren metrikak aurkezten dira. 5.2 azpiatalean esperimentuak burutzeko datuen antolamendua azaltzen da, zeintzuk izan diren erabilitako esaldi eta saioak zehaztuz. Bukatzeko, 5.3 azpiatalean esperimentuen emaitzak erakutsiko ditugu.

5.1 Ebaluazio metrikak

Ahots artifiziala sortzeko sistemetan, funtsezko bi ebaluazio-mota daude: ebaluazio objektiboa, emaitzetatik zuzenean lor daitekeen neurriren bat aplikatuz egiten dena; eta ebaluazio subjektiboa, lortutako audioaren kalitateari buruz pertsonei galdetzea eskatzen duena (adibidez, ulergarritasunari, naturaltasunari edo adierazkortasunari buruz galdetuz). Kasu honetan, aurretiazko esperimentuak direla kontuan hartuta, ebaluazio objektiboak bakarrik egitea erabaki da, ebaluazio subjektiboa, askoz ere neketsuagoa da, sistemak heldutasun handiagoa beharko luke eta, beraz, sintesi-kalitate handiagoa duenerako utziko da. Nolanahi ere, sortutako audioak proiektuaren webgunean [6] entzun daitezke.

Neurri objektibo posibleen artean, ereduaren galera-funtzioa kalkulatzeko erabili diren biak hautatu dira ebaluaziorako. Alde batetik, mel distantzia (MD), hau da, iragarritako mel espektrogramen eta *target*aren arteko distantzia euklideoa. Zenbat eta distantzia txikiagoa, orduan eta hobea izango da iragarpena. Bestalde, lortutako fonema-zehaztasuna (FZ) ([0-1] tarteko balioa) zenbat eta handiagoa izan, orduan eta sistema hobea izango da. Eredua probatzeko erabiltzen diren seinaleak isilean esandako

[6] https://aholab.ehu.eus/deeprestore/resultados/ (azken sartzea 25/11/2024).

esaldietatik datozenez eta xede-espektrograma entzuteko moduan esandako esaldietatik datorrenez, ereduak egindako iragarpenak DTWren bidez lerrokatzen dira neurketen balioa kalkulatu aurretik. FZ kalkulatzeko, fonemen iragarleak trama guztiekiko zuzen sailkatutako tramen proportzioa hartzen da kontuan.

5.2 Dato-multzoen prestaketa

3. atalean aipatzen den bezala, hiru saio mota daude: modu entzungarrian (*a*) soilik, modu isilean (*s*) soilik, eta modu entzungarrian nahiz isilean (*a+s* eta *a+2s*) grabatutako esaldiak. Experimentu hauetan isilean grabatutako esaldiak (*s*) dagokion saio bereko *a* seinalearekin DTW bidez lerrokatzen dira erreferentziazko ahots-seinalea lortzeko.

Esperimentuetan, datuak entrenamendu-, baliozkotze- eta proba-azpimultzoetan banatzen dira. Entrenamendu-azpimultzoetan *a*, *a+s* eta *a+2s* motatako adibideak erabili dira (ikus 1. taula). Baliozkotze- eta proba-azpimultzoek *saio misto*en *s* adibideak baino ez dituzte, ereduaren azken helburua pertsona laringektomizatuekin erabiltzea denez, modeloaren errendimendua grabazio isiletatik datozen EMG-seinaleekin baloratzea egokiena delako. Saioaren tamainaren arabera saio mistoen *s* esaldietatik 3-12 esaldi erreserbatu dira baliozkotze eta probarako, entrenamendu-azpimultzoan sartu ez direnak hain zuzen ere.

Azkenik, 006 eta 009 saioetarako *s* esaldiak soilik grabatu direnez ezin da audio erreferentziarik kalkulatu eta, ondorioz, ezin dira zuzen ebaluatu aurkeztutako ebaluazio metrikekin. Beraz ez dira erabili entrenamendurako, baliozkotzerako ezta probetarako.

Soilik modu entzungarrian grabatutako 4 saioak (001tik 004ra arte izendatuak) entrenamendu-azpimultzoetan baino ez dira erabili. Azkenik, 9 *saio misto* gelditzen dira ereduak entrenatzeko, balioztatzeko eta probatzeko erabili ahal direnak (005, 007, 008, 010-015).

Sistema SI baldintzetan probatzeko, 9 entrenamendu desberdinak egin dira. Entrenamendu bakoitzerako 001tik 004ra arteko saioak eta beste 8 *saio misto*en entrenamendu parteak erabili dira eredua entrenatzeko. Gero, SD baldintzetan balioztatzeko eta probatzeko 8 *saio misto* horien baliozkotze- eta proba-azpimultzoak erabiliko dira. Soberan geratzen den bederatzigarren *saio mistoaren* proba-azpimultzoa entrenatutako eredua SI baldintzetan probatzeko erabiltzen da.

3. taulan entrenamendu-, balioztatze- eta proba-azpimultzoen iraupenak erakusten dira. Iraupenen batezbestekoa 9 entrenamenduetarako sortutako azpimultzo desberdinak kontuan hartzen kalkulatu dira.

3. taula Datu-basearen iraupena. ReSSInt-001 sistemarako balioak (ReSSInt-EMG datu-basearen 001 esataria) esperimentuen batezbesteko balioak dira. *Trainset* entrenamendu-azpimultzoa da, *devset* baliozkotze-azpimultzoa da eta *testset* proba-azpimultzoa da, saioarekiko mendeko probak (SD) eta saioarekiko mendekotasunik gabekoak (SI) bereizten dituena.

	ReSSint-001	
Iraupena oo:mm:ss	**Audio**	**EMG**
Trainset	2:09:33 ± 0:03:46	3:29:16 ± 0:03:04
Devset	—	0:03:52 ± 0:00:15
Testset SD	—	0:07:13 ± 0:00:24
Testset SI	—	0:00:55 ± 0:00:25

5.3 Esperimentuak eta emaitzak

Datuen prestaketaren atalean esan bezala, sistema probatzeko, 9 eredu desberdinak entrenatu dira, entrenamendu bakoitzean 4+8 saio erabiliz eredua sortzeko (hau da, entrenamendu bakoitzean, *saio misto* bat kanpoan utzi da). Saioarekiko mendekotasunik gabeko baldintzetan (SI) ebaluazioa egiteko, eredu bakoitza kanpoan utzitako saiotik hartutako proba-esaldiekin probatzen da. Saioarekiko mendekotasunetako baldintzetan (SD) ebaluatzeko, ordea, eredua entrenatzeko eta balioztatzeko erabili diren saio mistoetatik hartutako proba-azpimultzoekin probatzen da. 4. taulak saio guztietarako emaitzen batezbestekoa erakusten ditu. Ikus daitekenez, SD baldintzetan eta SI baldintzetan lortutako emaitzak oso antzekoak dira. Honek esan nahi du sensoreen kokapena saioen artean mantentzeko erabili den estrategia arrakastatsua izan dela.

4. taula ReSSInt-EMG (ReSSInt-001) datu-baseko 001 esataria erabiliz lortutako proben emaitzen batezbestekoa eta desbiderapen estandarra azaltzen dira. MD: mel distantzia; FZ: fonema-zehaztasuna.

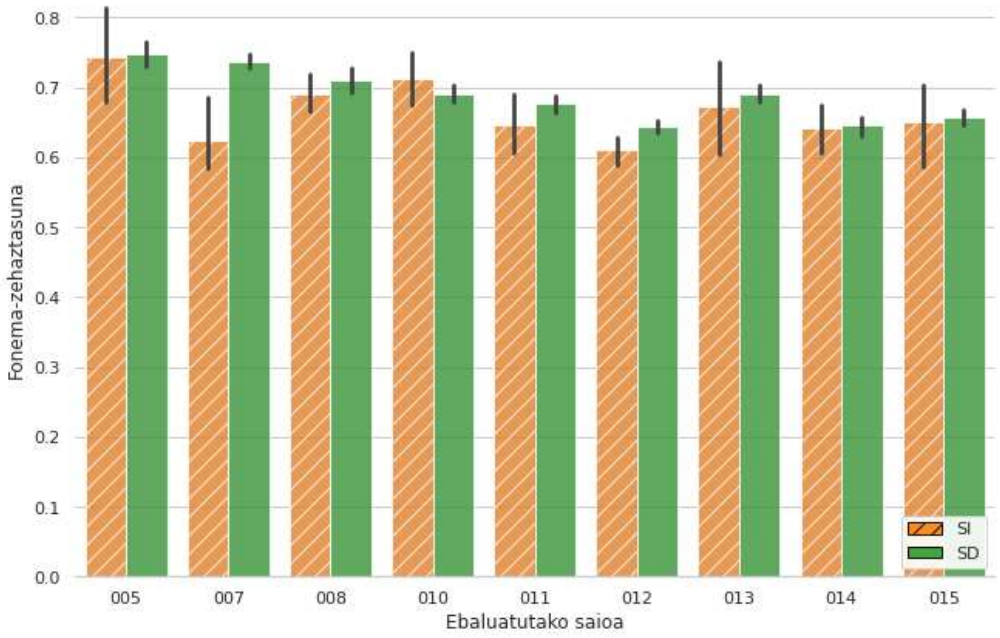
	MD	**FZ**
SD	2.929 ± 0.206	0.675 ± 0.065
SI	3.075 ± 0.230	0.662 ± 0.068

7. irudiak, 4. taulan erakutsitako datuak saiotan desagregatuta erakusten ditu (7.a irudia fonema-zehaztasuna, zenbat eta handiagoa izan, hobe; 7.b irudia mel distantzia, zenbat eta txikiagoa izan, hobe). Barrek deskribatutako bi baldintzak erakusten dituzte: SD eta SI (saioarekiko mendekotasuna edo mendekotasunik eza). Ardatza horizontalak erakusten duen identifikadoreak probatzen den saioaren identifikadorea da. SI emaitzen kasuan, identifikadorean adierazten den saioa entrenamendutik kanpo utzi den *saio mistoa* da, eta azaldutako emaitzak lortu dira *saio misto* hori gabe entrenatu den eredua probatuz. SD kasuan aldiz, saio horrekin entrenatu diren 8 ereduekin egindako proben emaitzen batezbestekoa da.

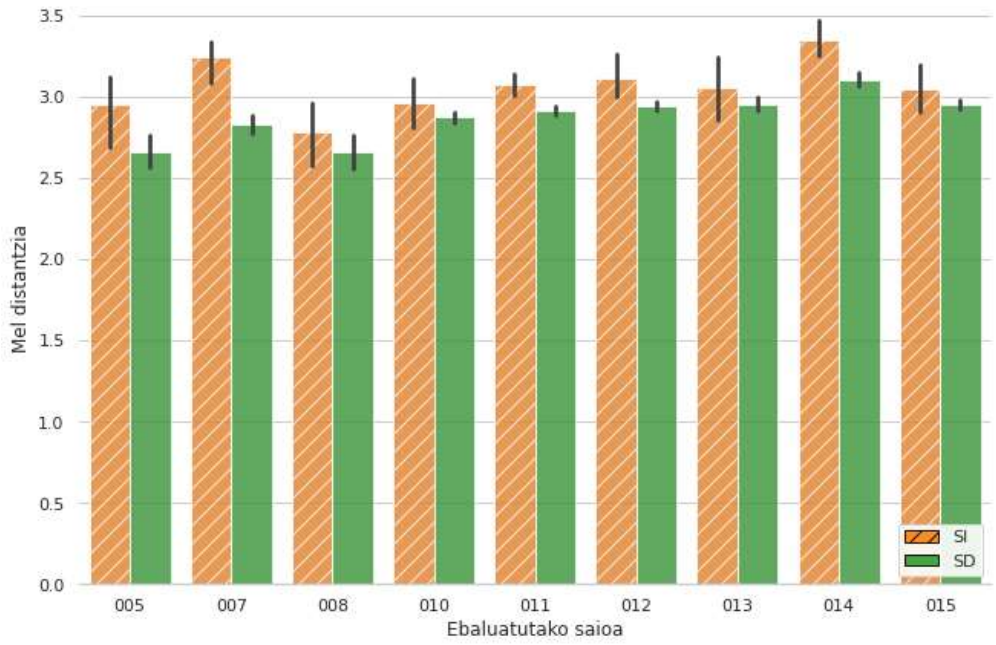
Ikus daitekeenez, orokorki fonema-zehaztasunaren emaitzak oso antzekoak dira bi baldintzetan, 007 saioan izan ezik. Saio horretan, badirudi emaitzak nabarmen okertu direla saioarekiko mendekotasunari dagokionez, bai mel distantziaren arabera, bai fonema-zehaztasunaren arabera. Bestalde, mel distantzia neurrian SI baldintzetan egindako probak behera egiten dutela hautematen da, desberdintasunak esanguratsuak ez diren arren (007 saioarako izan ezik).

Proiektuaren emaitzen web-orrian[7] atal honetan aurkeztutako esperimentuen lortutako audioak entzun daitezke.

[7] https://aholab.ehu.eus/deeprestore/resultados/ (azken sartzea 25/11/2024).



7.a irudia Fonema-zehaztasunarako lortutako emaitzen batezbestekoa. Zehaztasuna zenbat eta handiagoa izan, orduan eta hobea da sistema. Lerro beltzek neurriaren desbiderapen estandarra adierazten dute.



7.b irudia Mel distantziaren emaitzak. Mel distantzia zenbat eta txikiagoa izan, orduan sistema hobea da. Lerro beltzek neurriaren desbiderapen estandarra adierazten dute.

6 ONDORIOAK

Mendekotasunik gabeko saio-probetan lortutako errendimendua saioarekiko mendekotasuna duten probetan lortutakoa baino apur bat txikiagoa bada ere, aldea oso txikia da. Horrek adierazten du sentsoreak modu konsistentean kokatzeko 3D maskara erabiltzea eraginkorra izan dela saioen arteko aldakortasuna minimizatzeko. Ikuspegi praktiko batetik, ereduak entrenamenduan ikusi gabeko saioekin duen errendimendua eta lehendik ezagunak diren saioekin duena antzekoa bada, eredua benetako aplikazioetan erabil daiteke, erabiltzaileak sentsoreak jartzen dituen bakoitzean berrentrenatu edo egokitu beharrik gabe.

Ebaluazio subjektibo zehatzik egin ez bada ere, audioak modu informalean entzunez gero, ondorioztatzen da audio sintetikorako lortutako ulergarritasuna ez dela nahikoa. Gaztelaniazko hizketa-ezagutza sistema batekin egindako probek [41] % 64,7ko WER emaitzak eman zituzten, aplikazio errealista baterako nahikoa ez dena. Ingelesezko datu-basearekin egindako emaitzei erreparatzen badiegu [32], datu-base osoa erabiltzen ari denean (ia 17 orduko datuekin entrenatzen) egileak % 36,2ko WERa aipatzen du (proba azpimultzoan entrenamenduan erabilitako saioetako datuak sartzen direlarik, hau da SD baldintzetan). Bestalde, % 42,2ko WERa lortu da 4,6 orduko datu-base multilokutore batean

(txineraz eta beste sare-arkitektura bat erabiliz) [30]. Beraz, ondorio nagusia da datu gehiagorekin entrenatu behar dela emaitzak hobetu nahi badira.

7 ETORKIZUNERAKO LANAK

Lan honetan aurkezten diren emaitzak esatari bakar baten datuen ebaluazioari dagozkio. Esperimentu horiek frogatu dute EMG-seinaleetatik ahotsa sor daitekeela, baina datu gehiago bildu behar dira ulergarritasun onargarria lortzeko. Artikulu honetan aipatu ez diren arren, ReSSInt-EMG datu-baseak hiru esatari laringektomizatu eta bost esatari tipiko ditu, hautatutako esatariaz gain, baina datu kopuru txikiagoarekin.

Artikulu hau idazteko unean, gainerako esatari tipikoentzat eskuratutako datuak ebaluatzeko esperimentu-multzo bat aurreratuta dago. Atariko emaitzek erakusten dutenez, eredu unibertsal multilokutore bat egin daiteke, *speaker embedding* teknikekin egina. Teknika horien bidez, ereduari bektore bat ematen zaio, esatari bakoitzaren ezaugarriekin, gainerako sarrera-datuekin batera. Teknika horiek eredu bat esatari espezifiko batentzat doitzeko abiapuntu gisa erabil daitezke.

Berehala heldu nahi den arazo bat da hizketa-seinaleak sortzea laringektomizatutako esatarientzat, modu isilean baino ezin baitute artikulatu. Horretarako, *speaker embedding* bektoreak EMG-seinaleetatik abiatuta sortzeko estrategiak aztertzen ari gara, ahots-grabazioak erabili beharrean, ohikoa den bezala.

Aldi berean, sare-arkitektura berriak ikertzen ari gara, EMG-seinaletatik abiatuta audioaren iraupena zuzenean kalkulatzeko gai diren modulu neuronal konboluzionalak gehituz, [30] erreferentzian egindako lanari jarraituz.

Azkenik, aipatu nahi dugu ReSSInt-EMG datu-baseak audioa eta EMG-seinaleak aldi berean grabatutako esatarien aurpegiko irudiak ere badituela. DeepRestore proiektu berrian, ezpainak irakurtzeko modalitatea sartzea aztertzen ari gara. Espero dugu datu berriek seinaleen amaierako kalitatea hobetzea ahalbidetuko dutela, ereduak entrenatzeko behar diren datu-kopurua murriztuz.

8 ESKER ONAK

PID2022-141378OB-C21 proiektua, MCIN/AEI /10.13039/501100011033/-k, FEDER Una manera de hacer Europa-k eta Grant PID2019-108040RB-C21/AEI/10.13039/501100011033-k finantziatua.

BIBLIOGRAFIA

[1] DENBY, B., SCHULTZ, T., HONDA, K., HUEBER, T., GILBERT, J.M., BRUMBERG, J.S. 2010. "Silent speech interfaces" *Speech Commun.*, **52(4)**, 270-287.

[2] SCHULTZ, T., WAND, M., HUEBER, T., KRUSIENSKI, D. J., HERFF, C., BRUMBERG, J. S. 2017. "Biosignal-based spoken communication: A survey". *IEEE/ACM Trans. Audio Speech Lang. Process*, **25(12)**, 2257-2271.

[3] GONZALEZ-LOPEZ, J.A., GOMEZ-ALANIS, A., DOÑAS, J. M. M., PÉREZ-CÓRDOBA, J. L., GOMEZ, A. M. (2020. "Silent speech interfaces for speech restoration: A review" *IEEE Access*, **8**, 177995-178021.

[4] CHUNG, J. S., SENIOR, A., VINYALS, O., ZISSERMAN, A. 2017. "Lip reading sentences in the wild" *In Proc. IEEE CVPR*, 3444-3453.

[5] GONZALEZ, J. A., CHEAH, L. A., GOMEZ, A. M., GREEN, P. D., GILBERT, J. M., ELL, S. R., MOORE, R.K., HOLDSWORTH, E. 2017. "Direct speech reconstruction from articulatory sensor data by machine learning" *IEEE/ACM Trans. Audio Speech Lang. Process.*, **25(12)**, 2362-2374.

[6] ZHANG, Y., CAI, H., WU, J., XIE, L., XU, M., MING, D., YAN. Y., YIN, E. 2023. "EMG-based cross-subject silent speech recognition using conditional domain adversarial network" *IEEE Transactions on Cognitive and Developmental Systems*, **15(4)**, 2282 - 2290.

[7] LEE, K. 2022. "Ultrasonic Doppler Based Silent Speech Interface Using Perceptual Distance" *Applied Sciences* **12(2)**: 827. https://doi.org/10.3390/app12020827

[8] ANUMANCHIPALLI, G. K., CHARTIER, J., CHANG, E. F. 2019. "Speech synthesis from neural decoding of spoken sentences" *Nature*, **568(7753)**, 493.

[9] DASH, D., FERRARI, P., WANG, J. 2020. "Decoding imagined and spoken phrases from non-invasive neural (MEG) signals" *Frontiers in neuroscience*, **14**, 490970.

[10] CHAN, A. D., ENGLEHART, K., HUDGINS, B., LOVELY, D. F. 2001. "Myo-electric signals to augment speech recognition" *Medical and Biological Engineering and Computing*, **39**, 500-504.

[11] MAIER-HEIN, L., METZE, F., SCHULTZ, T., WAIBEL, A. 2005. "Session independent non-audible speech recognition using surface electromyography" *In IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*, 331-336.

[12] SCHULTZ, T., WAND, M. 2010. "Modeling coarticulation in EMG-based continuous speech recognition" *Speech Communication*, **52(4)**, 341-353.

[13] DIENER, L., JANKE, M., SCHULTZ, T. 2015. "Direct conversion from facial myoelectric signals to speech using deep neural networks" *In 2015 International Joint Conference on Neural Networks (IJCNN),* 1-7.

[14] MOSTAFA, S. S., AWAL, M. A., AHMAD, M.,RASHID, M. A. 2016. "Voiceless Bangla vowel recognition using sEMG signal" *SpringerPlus*, **5**, 1-15.

[15] SOON, M. W., ANUAR, M. I. H., ABIDIN, M. H. Z., AZAMAN, A. S., NOOR, N. M. 2017. "Speech recognition using facial sEMG". *In 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 1-5.

[16] MA, S., JIN, D., ZHANG, M., ZHANG, B., WANG, Y., LI, G., YANG, M. 2019. "Silent speech recognition based on surface electromyography". *In 2019 Chinese Automation Congress (CAC)*,4497-4501.

[17] WANG, Y., TANG, T., XU, Y., BAI, Y., YIN, L., LI, G., ZHANG, H., LIU, H., HUANG, Y. 2021. "All-weather, natural silent speech recognition via machine-learning-assisted tattoo-like electronics" *npj Flexible Electronics*, **5(1)**, 20.

[18] WU, J., ZHANG, Y., XIE, L., YAN, Y., ZHANG, X., LIU, S., AN, X., YIN, E., MING, D. 2022. "A novel silent speech recognition approach based on parallel inception convolutional neural network and Mel frequency spectral coefficient". *Frontiers in Neurorobotics*, **16**, 971446.

[19] LI, W., YUAN, J., ZHANG, L., CUI, J., WANG, X., LI, H. 2023. "sEMG-based technology for silent voice recognition" *Computers in Biology and Medicine*, **152**, 106336.

[20] MELTZNER, G. S., SROKA, J. J., HEATON, J. T., GILMORE, L. D., COLBY, G., ROY, S. H., CHEN, N., DE LUCA, C. J. 2008. "Speech recognition for vocalized and subvocal modes of production using surface EMG signals from the neck and face". *In INTERSPEECH,* 2667-2670.

[21] MELTZNER, G. S., HEATON, J. T., DENG, Y., DE LUCA, G., ROY, S. H., KLINE, J. C. (2017). "Silent speech recognition as an alternative communication device for persons with laryngectomy". *IEEE/ACM transactions on audio, speech, and language processing*, **25(12)**, 2386-2398.

[22] MELTZNER, G. S., HEATON, J. T., DENG, Y., DE LUCA, G., ROY, S. H., KLINE, J. C. 2018. "Development of sEMG sensors and algorithms for silent speech recognition" *Journal of neural engineering*, **15(4)**, 046031.

[23] GADDY, D., KLEIN, D. 2021. "An Improved Model for Voicing Silent Speech" *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,* **2**, 175–181.

[24] WAND, M., SCHULTE, C., JANKE, M., SCHULTZ, T. 2013. "Array-based electromyographic silent speech interface" *In International Conference on Bio-inspired Systems and Signal Processing* **2**, 89-96.

[25] DENG, Z., ZHANG, X., CHEN, X., CHEN, X., CHEN, X., YIN, E. 2023. "Silent speech recognition based on surface electromyography using a few electrode sites under the guidance from high-density electrode arrays" *IEEE Transactions on Instrumentation and Measurement*, **72**, 1-11.

[26] ZHU, M., ZHANG, H., WANG, X., WANG, X., YANG, Z., WANG, C., SAMUEL, O.W., CHEN, S., LI, G. 2021. "Towards optimizing electrode configurations for silent speech recognition based on high-density surface electromyography" *Journal of neural engineering*, **18(1)**, 016005.

[27] KUBICHECK, R. "Mel-cepstral distance measure for objective speech quality assessment." *Proceedings of IEEE pacific rim conference on communications computers and signal processing*. Vol. 1. IEEE, 1993.

[28] TOTH, A. R., WAND, M., SCHULTZ, T. 2009. "Synthesizing speech from electromyography using voice transformation techniques" *In Tenth Annual Conference of the International Speech Communication Association*, 652-655.

[29] JANKE, M., DIENER, L. 2017. "EMG-to-speech: Direct generation of speech from facial electromyographic signals" *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25(12)**, 2375-2385.

[30] LI, H., LIN, H., WANG, Y., WANG, H., ZHANG, M., GAO, H., AI, Q., LUO, Z., LI, G. 2022. "Sequence-to-sequence voice reconstruction for silent speech in a tonal language" *Brain Sciences*, **12(7)**, 818.

[31] SCHECK, K., SCHULTZ, T. 2023. "Multi-speaker speech synthesis from electromyographic signals by soft speech unit prediction" *In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*1-5.

[32] GADDY, D. 2022. Voicing Silent Speech. University of California, Berkeley.

[33] AUBANEL, V., LECUMBERRI, M. L. G., COOKE, M. 2014. "The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology". *International journal of audiology*, **53(9)**, 633-638.

[34] SAINZ, I., ERRO, D., NAVAS, E., HERNÁEZ, I., SANCHEZ, J., SARATXAGA, I., ODRIOZOLA, I. 2012. "Versatile Speech Databases for High Quality Synthesis for Basque". *In LREC*, 3308-3312.

[35] MCAULIFFE, M., SOCOLOF, M., MIHUC, S., WAGNER, M., SONDEREGGER, M. 2017. "Montreal forced aligner: Trainable text-speech alignment using kaldi". *In Proc. INTERSPEECH 2017*, 498-502.

[36] SALOMONS, I., DEL BLANCO, E., NAVAS, E. HERNAEZ, I. 2023. "Phone Confusion Analysis for EMG-Based Silent Speech Interfaces". *In Proc. INTERSPEECH 2023*, 1179-1183, doi: 10.21437/Interspeech.2023-1881.

[37] SALOMONS, I., DEL BLANCO, E., NAVAS, E. HERNAEZ, I., ZUAZO, X. 2023 "Frame-Based Phone Classification Using EMG Signals" *Appl. Sci*. 2023, 13(13), 7746.

[38] https://github.com/dgaddy/silent_speech (azken sarbidea 2024/09/17).

[39] 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.

[40] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, I., POLOSUKHIN, I. 2017. "Attention is all you need". *Advances in neural information processing systems*, **30**. https://doi.org/10.48550/arXiv.1706.03762

[41] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_es_conformer_ctc_large

IRUDIEN JATORRIA

1.a irudia: CNX Anatomy 2013, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>, via Wikimedia Commons. Artikulurako aldatuta. URL: https://upload.wikimedia.org/wikipedia/commons/c/c2/1106_Front_Views_of_the_Muscles_of_Facial_Expressions.jpg

1.b irudia: CNX Anatomy 2013, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>, via Wikimedia Commons. Artikulurako aldatuta. URL: https://commons.wikimedia.org/wiki/File:1106_Side_Views_of_the_Muscles_of_Facial_Expressions.jpg

1.c. irudia: Olek Remesz (wiki-pl: Orem, commons: Orem), CC BY-SA 2.5 <https://creativecommons.org/licenses/by-sa/2.5>, via Wikimedia Commons. Artikulurako aldatuta. URL: https://commons.wikimedia.org/wiki/File:Musculi_coli_base.svg