Hizketa-ezagutzan oinarritutako estrategiak, euskarazko online OBHI (Ordenagailu Bidezko Hizkuntza Ikaskuntza) sistemetarako

Igor Odriozola Sustaeta

Aholab Seinale Prozesaketako Laborategia
Komunikazioen Ingeniaritza Saila

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Aholkulariak: Inma Hernáez and Eva Navas

Euskal Herriko Unibertsitateari bidalitako tesia
Telekomunikazio Ingeniaritzako doktore-gradurako

March 13, 2019

Hizketa-ezagutzan oinarritutako estrategiak, euskarazko online OBHI (Ordenagailu Bidezko Hizkuntza Ikaskuntza) sistemetarako

Igor Odriozola Sustaeta

Aholab Seinale Prozesaketako Laborategia
Komunikazioen Ingeniaritza Saila

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Aholkulariak: Inma Hernáez and Eva Navas

Euskal Herriko Unibertsitateari bidalitako tesia
Telekomunikazio Ingeniaritzako doktore-gradurako

March 13, 2019

Lengoagetan ohi inçan
Estimatze gutitan
Oray aldiz hic behar duc
Ohoria orotan.

Heuscara
Habil mundu gucira

Bernart Etxepare, 1545

Tesi hau amari eskaini nahi diot, euskara bihotzetik erakutsi baitzidan, eta aitari, nigan zientziarekiko pasioa piztu baitzuen.

Laburpena

Ordenagailu Bidezko Hizkuntza Ikaskuntza (OBHI) sistemetan hizketa-teknologiak erabiltzeko interesa gero eta handiagoa da. Horren zergatia, hein handi batean, hizketa-teknologiek azken urteotan izan duten hobekuntza nabarmena da, eta gaur egun gero eta jende gehiagok erabiltzen ditu gero eta modu naturalagoan. Literaturak erakusten du OBHIn bi interesgune nagusi daudela hizketa-ezagutze automatikoarekin (ASR, *Automatic Speech Recognition*) loturik: Ordenagailu Bidezko Ebakera Lanketa (OBEL) eta Ahozko Gramatika Praktika (AGP). Ohiko OBEL aplikazio batean, esaldi bat irakurri eta grabatzeko eta ikasteko aplikazioaren zerbitzarira bidaltzeko eskatzen zaio erabiltzaileari. Zerbitzariak emaitza bat bueltatzen du, eskuarki hiru mailako kolore-sistema bat erabiliz, adierazten duena zer fonema ahoskatu den zuzen eta zein ez. AGP aplikazioak ez dira oraindik oso ezagunak, baina, besteak beste, aukera anitzeko testak daude, non erabiltzaileek ahoz hautatzen baitute erantzun bat hainbat aukeren artean; hizkuntzetarako tutoretza-sistema adimendunak ere badaude, non ikasleek ordenagailuak egindako galderei erantzun eta sistemak *feedback*a ematen baitie. Halako tresnak ikasleen ikasketa-prozesuarekiko autonomia indartzeko erabil daitezke, aukera ematen baitute ahotsa erabiliz ebakera hobetzeko edo ikasgelatik kanpo gramatika-ariketak egiteko.

Lan honetan, bi aplikazio kontsideratu dira: batetik, OBEL sistema klasikoa, non aurrez zehaztutako esaldi bat grabatzen baitu ikasleak, eta sistemak ebakerari buruzko *feedback*a ematen baitio; bestetik, Hitzez Hitzeko Esaldi Egiaztapenerako (HHEE) sistema berri bat, non esaldi bat sekuentzialki egiaztatzen baita, hitzez hitz, eta antzeman bezain laster erakusten baitzaio hitza erabiltzaileari. HHEE tresnarekin, gramatika-ariketak ahoz ebazteko tresna bat (AGP) eraiki daiteke. Sistema bien oinarrian, esakuntza egiaztatzeko teknikak daude, hala nola ebakera-egokitasun (GOP, *Goodness Of Pronunciation*) puntuazio ezaguna.

Sistema entrenatzeko hautatutako datu-base akustikoa *Basque Speecon-like* datubasea da, euskaraz publikoki erabilgarri dagoen bakarra, hizketa-ezagutzarako berariaz diseinatua eta mikrofono bidez grabatua. Datu-base horrek zenbait eragozpen ditu; adibidez, ebakera-lexikoia eta zenbait etiketa-fitxategi ditu faltan. Gainera, hizketa dialektal asko du, batez ere bat-bateko atalean. ASRn, aldaera fonetikoak eredu akustiko bakar batez modela daitezke; hala ere, OBEL sistemek eredu "garbiak" behar dituzte erreferentzia gisa erabil daitezen. Horrenbestez, zenbait lan egin behar izan da datu-

baseko etiketa-fitxategietan. Transkripzio asko aldatu egin behar izan dira, eta lexikoi berri bat sortu da ebakera dialektal alternatiboak kontuan hartuta. Azpimarratzekoa da lexikoi berriak4.12 ebakera desberdin dituela, batez beste, hitz bakoitzeko.

Tesi honetan hizketa ezagutzeko erabili den softwarea *AhoSR* da. *Aholab* ikerketa-taldean sortu eta garatu da, eta ezagutze-ataza desberdinekin jarduteko diseinatu da. Tesi honetako lana egin ahal izateko, esakuntza egiaztatzeko teknikak inplementatu zaizkio, oinarrizko atazekin batera exekutatzeko. Horretarako, grafo paralelo bat inplementatu da GOP puntuazioak kalkulatzeko. OBEL eta HHEE atazetarako, berariazko bilaketa-grafoak ere erantsi zaizkio, ataza bakoitzak berezko ezaugarriak ditu eta. Gainera, *socket*ak ere inplementatu dira *AhoSR*ko audio-sarreraren moduluan. Horrekin, denbora errealeko funtzionamendua lortzen da ezagutzailea Internet bidez atzitzean, eta horrek *AhoSR* zerbitzari batean instalatzeko aukera ematen du, atzipen unibertsala izateko.

Markoven ezkutuko ereduak (HMM, *Hidden Markov Models*) entrenatzeko modu desberdinak aztertu dira sakonki. Hasieran, kalitate hobeko HMMak espero genituen hiztegi alternatibadun berria erabilita. Hala ere, emaitzek ez dute hori erakusten, ebakera alternatiboen hain kopuru handiagatik. Eskuz zuzendutako datuak ere erantsi dira (entrenamenduko datuen % 15), alternatibak entrenamenduan hobeki lerrokatuko zirelakoan, baina sarrera bakarreko hiztegia erabiliz lortutako emaitzen antzekoak lortzen dira. Horrenbestez, eskuz zuzendutako transkripzioak ustiatzeko, HMMak entrenatzeko modu desberdinak aztertu dira. Hala, ikusi dugu ezen HMM pittin bat hobeak lortzen dira lehendabiziko etapetan transkripzio-akats gutxiko datuak eta, gero, datu-base osoa erabilita.

Hasierako sistema eraikitzeko, kontsideratu zen bi GOP banaketa behar zirela fonema bakoitza sailkatzeko: zuzen ebakitako fonemen banaketa eta oker ebakitakoen banaketa. Oker ebakitako fonemen GOPak lortzeko, transkripzioetan akats simulatuak txertatu ziren, eta transkripziook *AhoSR* deskodetzailetik pasarazi ziren lerrokatze behartu moduan. Ondoren, zuzen eta oker ebakitako fonemen atalaseak kalkulatu ziren, bi banaketen errore berdineko tasaren (EER, *Equal Error Rate*) puntu gisa. Metodo hori hasierako prototipo batean inplementatu zen, eta laborategiko zenbait esperimentu egin ziren, oso emaitza onak zian zituztenak. Orduan, ingurune errealistago batean testatu zen sistema: euskara-ikastegi edo euskaltegietan, 20 ikasleren artean. Emaitza objektiboak eta 20 ikasleek betetako inkestaren emaitzak benetan itxaropentsuak izan ziren.

Hasierako prototipoa lokalki exekutatzen zen, eta sistema unibertsalago baten beharra sumatu genuen, edozein gailutatik eta edonondik atzitzeko. Hala, HTML5 berriaren espezifikazioak baliatu genituen, nabigatzaileak, *audio API*aren bitartez, audio-sarrera atzi baitezake, plataforma edozein delarik ere. Gure sistema edozein sistema eragiletatik atzitu ahal izateko aukera eman digu horrek. Bestalde, HHEEan oinarritutako AGP atazarako, HTML5eko beste API bat erabili da (*web API*a), nabigatzailearen eta zerbitzariaren artean *socket* moduko konexioak sortzeko aukera ematen duena, audio-datuak grabatu ahala bidaltzeko.

Sistema *online* inplementatzeko, zenbait eragozpenekin egin dugu topo: adibidez, erabiltzaileek audioa jasotzeko erabiliko dituzten gailuen arteko desberdintasunak direla eta, nolabaiteko parametro-normalizazioa behar da. Are gehiago, *online* normalizazioko teknika bat behar da, HHEEn berehalako *feedback*a eman behar baita seinale osoa ezagutzailera iritsi aurretik. Hainbat teknika probatu dira batezbesteko- eta bariantza-normalizazio cepstrala (CMVN, *Cepstral Mean and Variance Normalisation*) inplementatzeko eta batezbestekoen eta bariantzen hasierako balioak estimatzeko. Emaitzarik onenak lan honetan proposatutako metodo hibridoaz lortu dira, non batezbestekoen hasierako balioak lehen N bilbeak erabiliz estimatzen diren eta bariantzen hasierako balioak entrenamenduko datu-multzotik erauzten diren.

Dena dela, CMVN teknika berri bat asmatu da tesi honetan: *normalizazio anitzeko puntuatzea* (MNS, *Multi Normalisation Scoring*) metodoan oinarritutako CMVN. MNSren funtsa da hainbat behaketa-egiantz puntuazio sortzea, sarrerako Mel maiztasuneko koefiziente cepstralak (MFCC, *Mel-Frequency Cepstral Coefficient*) baldintza desberdinetan grabatutako hizketa datu-multzo desberdinetatik kalkulatutako batezbestekoak eta bariantzak normalizatuz. MNSan oinarritutako CMVNa honetan datza: kalkulatzen da bilbe batek entrenamenduko datu-multzo bakoitzekoa izateko zer probabilitate duen; probabilitate horiek haztapen gisa erabil daitezke batezbestekoen eta bariantzen estimazioa kalkulatzeko. Lortu ditugun emaitzak nabarmentzekoak dira, batez ere seinale garbietarako. MNS erabiltzearen abantailarik handiena da CMVNa *online* erabil daitekeela, bilbez bilbe, inguruko bilbeak edo uneko bilbeari dagokion segmentuko bilbeak aztertu beharrik gabe.

MSN metodo bera erabiliz, *online*ko ahots-aktibitatearen detektagailu (VAD, *Voice Activity Detector*) berri eta eraginkor bat ere proposatu dugu. Baliozkotze-esperimentu batean, MNSan oinarritutako gure VADaren emaitzak eta ITU-Tko bi VAD algoritmo estandar (G.720.1 eta G.729b) alderatuz, emaitza orokor hobeak lortu ditugu, sailkapen-erroreak nabarmen txikiagoak baitira isiltasun-bilbeetarako, eta, antzekoak, berriz, hizketa-bilbeetarako. Gure sistema erabilgarria da, beraz, bai hizketan errore-tasa baxuak behar dituzten aplikazioetarako, bai isiltasunean errore-tasa baxuak behar dituztenetarako.

Azkenik, neurona-sareak (NN, *Neural Network*) erabili dira, sailkatzaile bat entrenatzeko unean parametro desberdinek duten eragina ikusteko asmoz. Horren ondorio gisa, ikusi dugu GOP puntuazioak direla parametrorik eraginkorrenak, aurreko, uneko eta ondorengo fonemen iraupenen eta log-egiantzen artean. Esperimentuen emaitzak koherenteak dira hasierako sistemaz lortutakoekin.

Eskertza

Eskerrak eman nahi dizkiet nire tesi-zuzendariei doktorego-tesi hau UPV/EHUko *Aholab* ikerketa-taldean egiteko aukera emateagatik, haien aholkularitza profesionalagatik, eta nigan jarritako konfiantzagatik.

Halaber, eskerrak eman nahi dizkiet ikerketa-taldeko nire lankide guztiei, beti laguntzeko prest daudelako, eta, bereziki, Luis Serrano nire lankideari, hainbeste alditan interes handiz lagundu izanagatik.

Eskerrak eman nahi dizkiet, orobat, tesi hau idazten jardun dudan bitartean nirekin egon diren pertsonei, eta honetan lanean ibili naizen denbora-tartean nire hutsunea sumatu dutenei.

Adierazpena

Hitz hauen bidez, adierazi nahi dut tesi hau nire lanaren eta ahaleginaren emaitza dela eta ez dela beste inora bidali sari baterako. Beste iturri eta informazio batzuk erabili ditudanean, hala direla adierazi da.

Laburdurak

AGP	Ahozko Gramatika Praktika (SGP, *Spoken Grammar Practice*)
ASR	*Automatic Speech Recognition* (Hizketa Ezagutze Automatikoa)
CMVN	*Cepstral Mean and Variance Normalisation* (Batezbesteko eta Bariantza Normalizazio Cepstrala)
FZ	Feedback Zuzentzailea (CF, *Corrective Feedback*)
GMM	*Gaussian Mixture Model* (Gaussiar Nahasteen Eredua)
GOP	*Goodness Of Pronunciation* (Ebakera Egokitasuna)
HHEE	Hitzez Hitzeko Esaldi Egiaztapena (WWSV, *Word-by-Word Sentence Verification*)
HMM	*Hidden Markov Model* (Markoven Ezkutuko Eredua)
L1, L2	*First language*, *second language* (Lehen hizkuntza edo ama-hizkuntza, bigarren hizkuntza edo xede-hizkuntza)
MFCC	*Mel Frequency Cepstral Coefficient* (Mel Maiztasuneko Koefiziente Cepstrala)
MLP	*Multi-Layer Perceptron* (Geruza Anitzeko Perzeptroia)
MNS	*Multi-Normalisation Scoring* (Normalizazio Anitzeko Puntuatzea)
NN	*Neural Network* (Neurona Sarea)
OBEL	Ordenagailu Bidezko Ebakera Lanketa (CAPT, *Computer Assisted Pronunciation Training*)
OBHI	Ordenagailu Bidezko Hizkuntza Ikaskuntza (CALL, *Computer Assisted Language Learning*)
PER	*Phone Error Rate* (Fonemen Errore Tasa)
VAD	*Voice Activity Detection* (Ahots Aktibitatea Detektatzea)

Contents

List of Figures

List of Tables

PART I

Sarrera

CHAPTER 1

Sarrera

1.1 Hitzaurrea

Oso zoriontsu sentitzen naiz tesi-lan hau egiteko aukera izan dudalako. Nire bi pasioak batzen ditu tesi honek: euskara eta zientzia. Ezin du jende askok esan maite duen lan batean diharduenik; nire kasuan, bikoiztuta gertatzen da hori.

Telekomunikazio-ingeniaritza ikaslea nintzela, euskara-eskolak ematen hasi nintzen, seguruenik bi alorrek erakarri izan nautelako beti, eta berehala jakin nuen nire etorkizun profesionalak bi alorrok batu beharko zituela nolabait. *Aholab*ek, hain ondo hartu nauen laborategiak, eman zidan hori egiteko aukera.

Laborategiak ikerketa-lerro berri bat zabaldu zuen haren interesguneen artean: bigarren hizkuntzak ikasteko software komertzialetan (euskararako), hizketa-ezagutze automatikoko (ASR, *Automatic Speech Recognition*) teknologia inplementatzea, ikaslearen eta ordenagailuaren arteko komunikazioaren kalitatea hobetzeko helburuaz, ikasleari laguntzeko bere ikasketa-prozesuari etekina ateratzen, autonomoki. Hura izan zen nire lehen proiektua *Aholab*en, eta halaxe hasi nintzen ASR eremuan lanean.

*Hizkuntzaren Azterketa eta Prozesamendua*n Masterra lortu ondoren, euskararako gure ASR sistema garatu nuen, gaur egun *AhoSR* izenez ezagutzen duguna. Hitz-gramatikan oinarritutako dekodetzailea zen, C++ hizkuntzan idatzia Windowserako, eta sarreratzat wav fitxategiak nahiz zuzeneko audioa onartzen zituen. Hastapenetako prototipo hura oinarri hartuta, *AhoSR* hazten eta hobetzen hasi zen.

Harrezkero zenbait urte pasatu dira, eta lehendabiziko ezagutzaile hura software osoago eta egonkorrago bat bihurtu da. Gaur egun, *AhoSR* beste ataza batzuetarako ere erabil daiteke, hala nola ezagutza fonetikoa, esakuntzaren egiaztapena, eta hiztegi handiko hizketa-ezagutze jarraitua (LVCSR, *Large Vocabulary Speech Recognition*). Web bidez atzi daiteke (HTML5 espezifikazioak baliatuz), erabiltzaileari urrunetik erabiltzeko aukera emanez, edozein dela ere erabiltzailearen plataforma. Zenbait demo proba daitezke hemen: http://aholab.ehu.eus/users/igor/demos.html.

1.2 Azalpen orokorra

Teknologiak aurrera egin ahala, hizkuntzen ikaskuntzan aplikatzeko aukera berriak sortzen dira. Azken urteotan agertu diren teknologiek eragin handia izan dute Ordenagailu Bidezko Hizkuntza Ikaskuntzaren (OBHI) alorrean, praktikan ez ezik, diskurtsoan, ikerketan eta pedagogian. Interneten hedapena (gaur egungo HTML5 web-zehaztapenak barne) eta mugikor-plataforma berriak (hala nola *smartphone*ak eta tabletak) iraultza globala izan dira, eta horrek ere eragin handia izan du hizkuntzak ikasteko eskolak bideratzeko eta antolatzeko moduan. Arloko aditu batzuek uste dute komunikatzeko eta, orobat, pentsatzeko modu berriak sorraraziko dituztela teknologia berri horien garapenak, eta, gainera, balitekeela hizkuntzak irakasteko arloan erabilitako pedagogia erabat berreraiki behar izatea [1].

OBHIaren alorrean, bai erabilerari bai ikerketari dagokienez, hauek dira, [2]en arabera, literaturan agertu diren elementurik erakargarrienak eta eragin handiena izan dutenak: ediziorako softwareak, ikasketa kudeatzeko sistemak, audio- eta bideo-konferentziako aplikazioak, adimen artifizialeko sistemak eta sistema adimendunak, teknologia mugikorrak, eta hizketa ezagutzeko eta ebakera lantzeko teknologiak. Horrenbestez, hizketa-teknologiak —bereziki, ASR eta Ordenagailu Bidezko Ebakera Lanketa (OBEL)— interespuntutzat jotzen dira gaur egungo OBHIaren arloan.

ASR-ren aplikazioak era askotakoak dira. Aplikazio horietariko bat da hizkuntzen tutoretza-sistema adimendunetan (ILTS, *Intelligent Language Tutoring Systems*) erabiltzea. Halako sistemetan, ordenagailuak emandako jarraibideei erantzun behar diete ikasleek; horretarako, haiek sortutako ahozko seinaleak jasotzen dira, erroreak hautematen dira, eta feedback-a ematen zaio ikasleari errore horien erantzun gisa. Hizketa-teknologien beste erabilera nagusi bat OBEL da. Ebakera lantzeko, alderdi segmentalak (bakarkako fonema-soinuak) nahiz supra-segmentalak (ezaugarri prosodikoak) kontsidera daitezke. OBEL teknologia inplementatzen duten aplikazio gehienetan, ikasleek sortutako soinuak eredu zuzen batekin alderatzen dira, eta ikusizko adierazpen baten bidez ematen da feedbacka. Teknologia horiek gero eta erabiliagoak dira beste eremu batzuetan ere, adibidez, hizketa patologikoaren detekzioan eta tratamenduan.

Web-teknologiaren aurrerapenen ondorioz, orain arte posible ez ziren interfaze berriak sortu dira. HTML5 zehaztapenak, *web audio API*aren bidez, aukera ugari ematen ditu Internet ahots bidezko aplikazioen bidez erabiltzeko. HTML5en bidez, ahots-sarrera bidezko web aplikazioak erraz konekta daitezke zerbitzarietan ostatatutako ASRekin eta ebakera lantzeko aplikazioekin, eta, hala, aukera gehiago sortzen dira bakarka ikasteko sistemetarako eta webean oinarritutako beste teknologia batzuetarako, hala nola ikasketa kudeatzeko sistemetarako eta bestelako ediziorako tresnetarako. HTML5ek, epe laburrean, Adobe Flash aplikazioak eta Java appletak ordezkatuko ditu, audioa berez kudeatzeko aukera ematen baitie web nabigatzaileei. Horrek esan nahi du HTML5 teknologia plataformarekiko independentea dela, nabigatzailearen araberakoa soilik. Gaur egun, nabigatzaile ezagunenek dagoeneko inplementatuta dute HTML5.

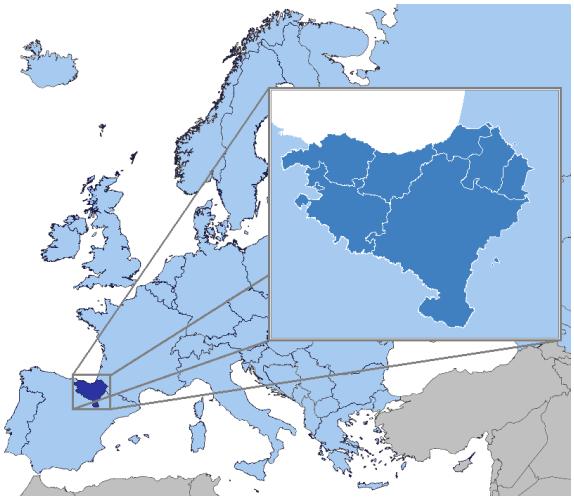
Tesi honetan, ASRan oinarritutako euskararako OBHI sistema bat eta OBEL sistema

bat inplementatzeko estrategia azaltzen da, ASR-rako datu-base estandar bat baliatuz. Sistema Interneten on-line inplementatzeko aurkitutako arazoak ere azaltzen dira, baita horiek gainditzeko proposatutako bideak ere.

1.2.1 Tesiaren motibazioa

Euskara Europako hizkuntza isolatu bat da, eta uste da gaur egun arte iraun duten hizkuntza aurreindoeuropar gutxietariko bat dela eta bakarra, mendebaldeko Europan. Hizkuntzaren jatorria ez da zehatz-mehatz ezagutzen, baina gaur egungo teoriarik hedatuenak dio proto-euskara edo euskararen forma zaharrago bat bazela hizkuntza indoeuroparrak iritsi baino lehen; hau da, euskararen lurraldea edo Euskal Herria geografikoki inguratzen duten hizkuntza erromantzeak.

Euskal Herria bi zatitan banatuta dago administratiboki: Ipar Euskal Herria, Frantziako administrazioaren menpe dagoena, eta Hego Euskal Herria, Espainiako administrazioaren menpe dagoena. Ipar Euskal Herriak Pirinio Atlantikoak departamenduaren ia erdia osatzen du, eta euskarak ez du onarpen publikorik han; izan ere, Frantziar Errepublikan hizkuntza ofizial bakarra dago: frantsesa. Hego Euskal Herria ere bitan zatitua dago: Euskal Autonomia Erkidegoa (EAE) eta Nafarroa (garai bateko Nafarroako Erresuma). Euskara EAEn ofiziala da espainierarekin batera; Nafarroan, berriz, estatus heterogeneoa du, eremuaren araberakoa. Euskara, beraz, diglosia-egoeran dago, hizkuntza ofizialek, ez-ofizialekin alderatuta, prestigio-estatusa baitute. Hori kontuan harturik, espainierak eta frantsesak eragin handia dute euskaran. Eusko Jaurlaritzaren 2012ko inkesta baten arabera, Euskal Herriko biztanleen % 27k hitz egiten du euskara (2 648 998tik 714 136k). Hiztun horietatik 663 036 Hego Euskal Herrikoak dira; gainerako 51 100 hiztunak, berriz, Ipar Euskal Herrikoak [3].



(Wikipediatik hartutako irudia)

Fig. 1.1: Euskal Herriaren kokapena Europan.

Zenbakiek agerian uzten dutenez, euskararen erabilera urruti dago normalizaziotik. Gaur egun jende asko dabil euskara ikasten. Udaletako eta erkidegoetako administrazioek diru-laguntzak ematen dizkiete, zenbait baldintza betez gero, euskara ikasi nahi dutenei. Euskararen irakaskuntzak bilakaera handia izan du azken 40 urteotan, eta Eusko Jaurlaritzako HABE erakundeak kudeatzen ditu hizkuntza-eskolak. Hizkuntza erabilienen mailako aukera teknologikoak merezi ditu euskararen irakaskuntzak ere.

Euskarak ahulgune bat du. Euskara batu edo estandarra ez zen sortu 1960ko

hamarkadaren amaierara arte, eta, horren ondorioz, euskara batua ez da oraindik inoren berezko hizkuntza "erreala". Gaur egun, hezkuntzako maila guztietan erabiltzen da, lehen hezkuntzatik unibertsitateraino, bai telebistan bai irratian, baita idatzizko ekoizpen gehienetan ere [4]. Dena dela, idatzizko asmoz sortu zen, batez ere, euskara. Horren adierazgarri garbia da Euskaltzaindiak euskara batuaren ebakeraren eta prosodiaren bere lehen lan sakona [5] 2014ra arte kaleratu ez izana, alegia, lehen urratsak eman eta 50 urtera. Denbora-tarte horretan, euskara-irakasleak erabat zehaztu gabeko hizkuntza bat irakasten aritu dira, eta ikasleek, euskara batuaren benetako hiztunek, forma eman behar izan diote ahozko euskara batuari.

Arestian esan bezala, euskara batua ez da oraindik inoren berezko hizkuntza "erreala". Euskaldun zaharrek beren aldaera dialektala erabiltzen dute, eta euskara batua eskoletan irakasten bada ere, oso bertsio formaltzat eta zenbaitetan artifizialtzat ere hartzen da. Gainera, euskarak izugarrizko aniztasun dialektikoa du. Historian zehar hainbat saiakera egin dituzte euskal hizkuntzalariek Euskal Herria euskalki eta azpi-euskalki geografikotan sailkatzeko. Hala ere, ia herri bakoitzak du bere aldaera [6]. Horrek agerian uzten ditu euskara-irakasleek gainditu behar dituzten zailtasunak ikasgelan euskara irakasteko.

Euskararen normalizazio-egoera honetan, OBHI sistemetan hizketa-teknologiak edukitzea laguntza handikoa izan daiteke. Tesi honetan proposatutako tresnak oso lagungarriak izan daitezke euskara-ikasleentzat, euskaldun zaharren moduan hitz egiteko eta kalitatezko ahozko euskara bat lortzeko bidean. Izan ere, ahotsa erabiltzeko aukera dute, ebakera ebaluatzeko ez ezik, gramatika-ariketak ebazteko ere, eta hori lagungarria izan daiteke ikasleen ahozko ekoizpen-trebetasunetarako. Gainera, ikasleak motibatuago ere senti daitezke halako tresnekin, batez ere bigarren hizkuntza ikasgelatik kanpo praktikatzeko aukerarik ez duten ikasleak.

1.2.2 Helburuak

Tesi honen helburu orokorra da ASRan oinarritutako estrategiak aztertzea, euskara ikasteko OBHI aplikazioetan inplementatzeko. Horretarako, esakuntza egiaztatzeko teknikak kontsideratu dira, ezagutzako deskodetzailean zehar kalkulatzen direnak. Teknikok bi abantaila eskaintzen dituzte, besteak beste: alde batetik, esakuntza baten ebakera ebaluatzeko aukera (OBEL sistemak), *Basque Speecon-like* datu-base akustikoko hizketa erreferentziatzat hartuta. Beste aldetik, erabiltzailearen erantzuna hitzez hitz eta denbora errealean egiaztatzen duen tresna bat izateko aukera, ariketak ahoz eginez gramatika bakarka praktikatzeko lagungarriak izan daitezkeenak. Teknika horri Hitzez Hitzeko Esaldi Egiaztapena (HHEE) deitu diogu lan honetan, eta hemendik aurrera termino hori erabiliko da teknika horri erreferentzia egiteko.

Helburu horretara iristeko eta ateratako konklusioen erabilgarritasuna bermatzeko, honako helburu partzial hauek ere izan dira kontuan:

- ASRak garatzeko berariaz diseinatutako datu-base batetik kalitate oneko eredu akustikoak sortzea.

- Konfiantza Puntuazioetarako erabaki-atalaseak garatzea eta hobetzea. Horretarako, fonemak taldekatzeko modu desberdinak ere aztertu behar dira.
- Halako sistema bat *on-line* inplementatzeko sortzen diren eragozpenak aztertzea, baita diseinu-konponbide bat proposatzea ere.
- Sistemaren funtzionamendu orokorra hobetzeko seinale-prozesaketako beste teknika batzuk aztertzea eta aplikatzea, ahal nola batezbesteko- eta bariantza-normalizazio cepstrala (CMVN, *Cepstral Mean and Variance Normalisation*) eta ahots-aktibitatea detektatzea (VAD, *Voice Activity Detection*).

1.2.3 Tesiaren egitura

Tesi hau 4 ataletan banatuta dago, eta 11 kapitulu ditu.

I. atala sarrera da, eta 2 kapituluz osaturik dago. Lehen kapitulua uneko hau da (1. kapitulua). Bigarren kapituluan (2. kapitulua), hizketa-teknologien eremuan OBHI sistemetarako argitaratu den literaturaren berri ematen da. Arreta jarriko dugu halako aplikazioetan erabiltzen diren ASR teknologietan, baita ebakera puntuatzeko erabiltzen diren teknika desberdinetan ere. Azkenik, zenbait ondorio azaltzen dira.

II. atalak hasierako sistemaren oinarriak aurkezten ditu. 3. kapituluan, tesi honetan erabilitako datu-base akustikoa eta hari kalitate oneko eredu akustikoak lortzeko egin behar izandako moldaketak azaltzen dira. Datu-basearen deskripzio xehea ageri da han, baita datu-basean egindako zenbait lanketa etiketen fitxategiak hobetzeko eta lexikoia sortzeko.

4. kapituluan, AhoSR-ren egitura eta funtzionaltasunak azaltzen dira. AhoSR tesi honetan oinarri gisa erabili den hizketa ezagutzeko softwarea da, *Aholab* ikerketa-taldeak sortua eta garatua. Kapitulu horretan, AhoSRn egindako aldaketak eta moldaketak ere ageri dira, egiaztapen-puntuazioen eta denbora errealeko hitzez hitzeko egiaztapenaren erabilera inplementatzeko.

5. kapituluan, eredu akustikoak entrenatzeko modu desberdinen analisi xehea azaltzen da. Tesi honetan erabilitako datu-baseak zenbait eragozpen ditu; izan ere, aldaera dialektal ugari ditu. Deskribatzen da nola sortu den lexikoia, aldaera dialektalak alternatiba gisa txertatuta; baita lexikoi hori erabiltzeak ere zer eragin dituen Markoven ezkutuko ereduak (HMM, *Hidden Markov Model*) entrenatzeko unean. Gainera, azaltzen da entrenamenduan azpimultzo desberdinak erabiltzeak zer ondorio dituen HMMetan, baita entrenamendu-prozesuko fase desberdinetan erabiltzeak ere. Hala lortutako HMMen ezagutze fonetikoko emaitzak azaltzen dira, eta azkenik zenbait konklusio aurkezten dira.

Ebakera-atalaseak ezartzearen desabantailetariko bat da oker ebakitako datuen gabezia. 6. kapituluan azaltzen da tesi honetan zer prozedura erabili den gai horri aurre egiteko, baita hori baliozkotzeko zenbait esperimentu ere. Atalaseak lortzeko prozedura hori ingurune errealista batean ebaluatzeko sortutako aplikazioa ere aurkezten da hemen. Aplikazio hori gramatika-ariketak ahoz ebazteko diseinaturik dago (HHEE sistema erabiliz) eta modu lokalean exekutatzen da ordenagailu bakoitzean. Azkenik, benetako

ikasleekin egindako ebaluazioari buruzko xehetasunak eta ebaluazio horretan lortutako emaitzak azaltzen dira.

III. atalean, hasierako sistema oinarri hartuz garatutako zenbait hobekuntza deskribatzen dira. 7. kapituluan, bezero/zerbitzari inplementazio-kontuak deskribatzen dira. Azken urteotan garatu den HTML5 zehaztapenetan, bi funtzionaltasunak baliatuko dira: *web audio API*a eta *websocket API*a. Lehena audioa nabigatzailearen bidez grabatzeko erabiltzen da; bigarrena, berriz, audio-datuak *Nodejs* zerbitzari batera bidali eta feedbacka eskuratzeko.

8. kapituluan, audio-fitxategietan hizketa-segmentuak hautemateko metodo berri bat aurkezten da, normalizazio anitzeko puntuatzea (MNS, *Multi-Normalisation Scoring*) oinarri duena. MNS behaketa-egiantzetan oinarritzen da, alegia, normalizazio cepstrala aplikatuz entrenatutako isiltasun-HMMaren erdiko egoerako gaussiar nahasteen ereduan (GMM, *Gaussian Mixture Model*) sortzen diren behaketa-egiantzetan. Esperimentu bat egin da teknika hori erabiliz eta beste sistema ezagun batzuk erabiliz lortzen diren emaitzak alderatzeko, eta ondorioztatu dugu emaitzak oso lehiakorrak direla zarata-maila desberdinetan.

7. kapituluan, kalkulatu ahala cepstrumak normalizatzeko beste teknika berri bat aurkezten da, hori ere MNS teknikan oinarritua. Cepstrumen normalizazioa da HHEE sistema bat zerbitzari batean inplementatzean sortzen den arazoetariko bat. Denbora errealeko normalizazio cepstrala kudeatzeko estrategia desberdinak azaltzen dira, eta aterako konklusioak ere bai.

In 10. kapituluan, oker ebakitako fonemaren kontzeptua berraztertzen da. Horretan oinarriturik eta OBEL sisteman sailkatzaile gisa erabiltzeko, Neurona Sare desberdinak entrenatu dira, parametro multzo desberdinak erabiliz, asmoa baita parametro bakoitzaren eragina ikustea. Esperimentuen emaitzak eta zenbait konklusio azaltzen dira amaieran.

IV. atalak laburpena eta deskripzio orokorra azaltzen ditu. Azkenik, 11. kapituluan, tesi honetatik ateratako konklusio orokorrak eta etorkizuneko ikerketa-lanari buruzko zenbait hausnarketa ageri dira.

CHAPTER 2

ASR teknologia OBHI sistemetan

2.1 Sarrera

Ordenagailu Bidezko Hizkuntza Ikaskuntza (OBHI) hizkuntzalaritza aplikatuko eremu espezializatu bat da, hizkuntzen ikaskuntzako eta irakaskuntzako teknologiaren erabilerarekin zerikusia duena. Hastapenetan, hirurogeiko hamarkadan zehar alegia, errepikapen-ariketetara bideratutako softwareetan oinarriturik zegoen batez ere, behaviorismoaren eta hizkuntzak irakasteko ikuspegi audiolingualaren markoaren barnean. Hala ere, XX. mendearen amaieran ordenagailu pertsonalak eta multimedia eskuragarriagoa eta erabilgarriagoa bihurtu ahala eta Internet hedatu ahala, OBHI dibertsifikatu, zabaldu eta garatu egin zen [7].

OBHI, hastapenetan, ingeles-hizkuntzaren irakaspenaren (ELT, *English Language Teaching*) mendeko eremutzat jotzen zen [8], eta bigarren hizkuntzaren eskurapenari eta pedagogiari buruzko liburuek leku gutxi eskaintzen zioten teknologiaren erabilerari [9]. Gaur egun, OBHIaren alorrean teknologiaren erabilera sustatzen duten zenbait erakunde profesional daude —adibidez, *CALICO*, *EUROCALL* eta *IALLT*—, bai eta berariaz alor horretan diharduten aldizkari espezializatuak ere —adibidez, *Language Learning & Technology*, *ReCALL*, *CALL*, *Journal of Computer-Mediated Communication* eta *CALICO Journal*—. OBHI, beraz, ondo errotutako alorra da, ikerketa-agenda zabala duena eta askotariko aplikazio praktikoak dituena bigarren hizkuntzaren eskurapenaren alor guztietan barrena.

Praktika dago OBHIren bihotzean, eta ordenagailuak erabili izan dira hizkuntzen ikaskuntzako eremu guztietan. Ordenagailuok bereziki erabilgarriak kontsideratu izan dira gramatika, hiztegia, irakurmena eta idazmena, entzumena eta ebakera lantzeko, eta oso erabiliak izan dira hizkuntza ebaluatzeko. Autore askok behar-beharrezko elementutzat dute teknologia, bai ikasgela barnean, bai kanpoan, nahiz eta betiere pedagogiaren eta hizkuntza-eskurapenaren printzipioak kontuan dituen marko baten barnean. Ikasgelatik kanpo irakasleak teknologiaren erabilera modu egokian ikuskatuz gero, litekeena da ikasleek aprobetxamendu hobea izatea [10].

Gaur egun, autoikaskuntzako gelek bigarren hizkuntzak autonomoki ikasteko aukera

berriak eskaintzen dituzte. Hezkuntzako teknologiaz hornitu dira gehienbat gelok, eta, horrenbestez, autoikaskuntza eta teknologian oinarritutako ikaskuntza sinonimo bilakatu dira. Autonomia gai garrantzitsu bilakatu da, batez ere OBHI alorrean. Ikaskuntza autonomoaren abantailak, desabantailak eta ondorioak sakonki aztertzen ari dira gaur egun. Dena dela, autoikaskuntzaren kasuan bezala, autonomian ikerketan dihardutenek nabarmentzen dute teknologian oinarritutako ikaskuntzara jotzen duten ikasleek ez dutela zertan autonomoago izan teknologia erabiltzeagatik soilik; neurri handi batean, teknologiaren izaeraren araberakoa eta ematen zaion erabileraren araberakoa da hori [11]. Ikuspuntu bera ageri da [12]n, non azaltzen baita ezen, *L2* ikaskuntzan teknologia txertatzeko, ikasleen eta irakasleen beharrizan desberdinetara moldatu behar dela. Halaber, azaltzen da teknologia esanahia kudeatzeko eta ikasteko ere bideratu beharko litzatekeela, ez bakarrik ariketa baten ostean puntuazio bat emateko.

Gaur egun, OBHI alorrean, multimedia-software espezializatu eta sofistikatuak ez ezik, baliabide ugari erabiltzen dira, hala nola web-baliabideak, Web 2.0 tresnak eta sare-sozialetako softwareak, ikasketa kudeatzeko sistemak eta irakaskuntza-tresnak, eta mugikor-teknologiak. Baliabide horiek hainbat mailatan erabiltzen dira hizkuntzak ikasteko eta irakasteko asmoz, bai ikasgelan bertan, bai kanpoan. OBHIren aplikazio-aniztasun horren ondorioz, hizkuntzak ikasteko ikasgelan termino ugari sortu dira tresna teknologikoen erabilera izendatzeko. Hona, batzuk:

- Sarean oinarritutako hizkuntza-irakaskuntza (NBLT, *Network-Based Language Teaching*) [13]
- Ordenagailu bitartezko komunikazioa (CMC, *Computer Mediated Communication*) [14]
- Webaz indartutako hizkuntza-ikaskuntza (WELL, *Web-Enhanced Language Learning*) [15]
- Mugikor bidezko hizkuntza-ikaskuntza (MALL, *Mobile Assisted Language Learning*) [16]
- OBHI adimenduna (ICALL, *Intelligent CALL*) [17]
- Ordenagailu bidezko hizkuntza-ebaluazioa (CALT, *Computer Assisted Language Testing*) [18]
- Ikaskuntzarako teknologia (eLearning, *Learning with Technology*) [19]

Nahiz eta OBHI aplikazioak askotarikoak izan eta akronimo zerrenda luzea egon, 'OBHI' terminoak jarraitzen du hizkuntzen irakaskuntzan eta ikerketan teknologiaren erabilera adierazten duen aterki-termino gisa [1]. Hala eta guztiz ere, gaur egun OBHI terminoa OBHIA edo OBHI Adimendun terminorantz eboluzionatuz doa. OBHIA terminoak Hizkuntza Naturalaren Prozesamendua (HNP) eta Adimen Artifizial (AA) bidezko modelatzea integratzen ditu OBHIn, ordenagailuen eta erabiltzaileen arteko interakzioa hobetzeko eta ikasketa-esperientzia ikasle bakoitzari egokitzeko asmoz.

2.2 ASR teknologia OBHI sistemetan

1996an, [20] lanaren egileak bere azterketa batetik ondorioztatu zuen ordenagailuaren potentzial osoa baliatzen duten OBHI programek soilik emango zituztela ikasketa-emaitza hobeak, batez ere *feedback* egoki eta berehalako baten bidez. Gaur egun, esan genezake ezen, ASR teknologiak izan dituen aurrerapen nabarmenei esker, ordenagailuaren eta erabiltzaileen artean komunikazio naturalistago bat lortzetik hurbilago gaudela eta *feedback* egokiagoak ematen direla.

OBHI sistemek idatzizko ekoizpena lantzeko balio dute batez ere, baina OBHIk onura izan ditzake ahozkoaren praktikan ere [21]. Praktika garrantzitsua da, ahozko ekoizpenak karga kognitibo eta artikulazio-sistemaren kontrol handiagoak eskatzen ditu eta [22]. Karga kognitiboaren eta artikulazioaren kontrola entrenatzeko, OBHI sistemek ahozkoa praktikatzeko aukera eman eta hizketa-jardueraren Feedback Zuzentzaile (FZ) automatikoa eskaini beharko lukete. [23] artikuluan, OBHIren eraginkortasunari buruzko hausnarketa bat aurkezten da, eta egileak dio sakonago ikertu beharra dagoela OBHIk "*online*ko mintzamena"ren eremu "ia aztertu gabe"an duen eragina.

ASR teknologiaren alorreko aurrerapenek ahozko ekoizpenean oinarritutako ariketak garatzeko aukerak sorrarazi dituzte [24]. 350 ikaslerekin OBHI sistemen eraginkortasunaren berri emateko egindako ikerketa batean [25], Golonka et al.-ek frogatzen dute OBHI teknologia erabilgarria izan daitekeela ebakera lantzeko (Ordenagailu Bidezko Ebakera Lanketa, OBEL) eta ASR teknologia erabiltzen dela ahozkoa praktikatzeko, nahiz eta ez den aipatzen Ahozko Gramatika Praktikan (AGP) FZ automatikoa inplementaturik duen sistemarik. ASR darabilten OBHI sistemen berrikuspen-lan batean [26], ageri da urte hartan (2011) bazirela trebetasun komunikatiboak edo ebakera lantzeko sistemak, baina ez zegoela ahozko praktikako sistemarik gramatika-erroreen FZ automatikoa eskaintzen zuenik. Hala ere, gramatika hizkuntza-gaitasunaren alde garrantzitsua denez eta, hortaz, *L*2 ikaskuntzako helburu pedagogiko nagusietariko bat, zentzuzkoa da pentsatzea gramatika ahozko jardunean praktikatzeko sistema bat oso erabilgarria izan daitekeela *L*2 ikaskuntzarako.

XX. mendearen amaieratik hona, ASR teknologian oinarritutako hainbat sistema garatu dira, *L*2 ikaskuntzaren praktika eta FZa eskaintzen dutenak. Gehienek OBEL dute, batzuek ahozkoaren praktika dute inplementaturik, eta baten batek AGP darabil. Hona, zenbait adibide: *FLUENCY* [27], *Tactical Language Training System* [28], *AzAR* sistema [29], *SPELL* sistema [30][31], *Carnegie Speech Native Accent* [32], *Saybot* [33], *Euronounce* sistema [34], *EduSpeak* [35] eta *Tell me More* 2013an *Rosetta Stone*k erosia (`www.rosettastone.com`, 2018). Gaur egun, mugikorretarako eta tauletarako app ugari daude, hala nola *Babble*, *Busuu*, *Mondly* eta *Rocket Languages*. Hizkuntzak ikasteko app-en mugimendu handia dago orain merkatuan.

Datozen ataletan, xeheago deskribatzen da AGP eta OBEL alorren artearen egoera. AGP eta OBEL ASRan oinarritutako bi inplementazio dira (ikus 2.1. irudia), tesi honetarako hautatu direnak. Azken urteotan OBEL alorra landu duten ikerketa eta argitalpen asko izan badira ere, badirudi ikerketa eta baliabide gehiago behar direla

AGP alorrerako.

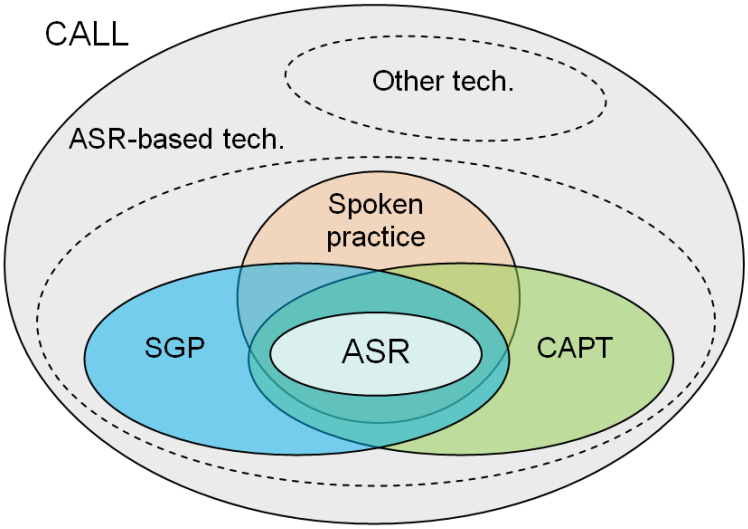


Figure 2.1: ASR teknologiak AGP eta OBEL barnean duen lekua OBHI alorrean, erabilitako teknologia motaren arabera.

2.2.1 Ordenagailu Bidezko Ebakera Lanketa (OBEL)

Ebakera-akatsak fonema- eta prosodia-akatsetan banatu daitezke [36]. Fonema-akatsei dagokienez, $L2$ ikasleek egin ditzaketen akatsik garrantzitsuenak hauek dira: ordezkapenak, ezabaketak eta txertaketak. Garrantzi gutxiagoko akastzat hartzen dira ebaki bai baina, jatorrizko hiztun batek ebakiko lukeenarekin alderatuta, soinu aski desberdina duten fonemak, eskuarki hiztunak doinu berezi bat duela erakusten dutenak. Horiek ere ordezkapen gisa ikus daitezke, baina ez dute komunikazioa oztopatzen. Prosodia-akatsei dagokienez, doinu ez-jatorrizkoa azentuaren, erritmoaren eta intonazioaren baitan sailkatu daiteke. Akats horiek lotura estua dute bata bestearekiko. Horrenbestez, dimentsio anitzeko auzia da ebakerarena, eta zaila da ikuspegi batez bakarrik ebaztea.

a) Fonema-akatsak

Ebakera automatikoki puntuatzeko lehendabiziko lanak 90eko hamarkadaren hasieran argitaratu ziren. Batez ere hitz mailan eta esaldi mailan diseinatuta zeuden, intonazio-, azentu- eta erritmo-neurriez osatuak: [37] lanean deskribatzen den sistemak eta *SPELL* (Interactive System for Spoken European Language Training) proiektuan garatutako prototipoak [38] seinale-prozesaketako teknikak baliatzen dituzte —hala nola antzekotasun-neurriak, distantzia espektrala eta funtsezko harmonikoaren eta energiaren arteko desberdintasunak— ikasle batek esandako hitz edo esaldi baten ebakera-kalitatea neurtzeko. [39] artikuluan, HMMak erabili zituzten nederlanderazko hitzak ebaluatzeko, ezagutze behartua baliatuz eta jatorrizko eta ez-jatorrizko hiztunek balioetsitako datuak

erabiliz. Hala ere, halako sistemetan jatorrizko hiztunen grabazio asko behar dira, material didaktikoko hitz bakoitzarentzat eredu bat entrenatu ahal izateko. Beraz, testuaren araberakok ziren, eta horrek esan nahi du hitz berrien ereduak sortzeko grabazio berriak egin behar direla.

Urte haietan, HMMetan oinarritutako ASRa, hizketa-unitate txikiagoak bainoago, esaldi osoak puntuatzeko ere erabiltzen zen [40][41]. Zenbait artikulutan, fonema-akats jakin batzuk irakasteko helburua zuten sistemak ere deskribatzen dira: [42]an, Viterbi deskodetzailearen iraupen-informazioa baliatu zen. [43]an, ebakera okerraren puntuazioa (MP, *mispronunciation score*) erabili zuten, zeina jatorrizko eta ez-jatorrizko hizketen egiantzen arteko ratioa baita. [44]an, HMMan oinarritutako hiru neurri alderatzen dira: log-egiantza, ondorengo log-probabilitatea eta segmentuaren iraupenaren puntuazioa; hiruren artean, antza, ondorengo probabilitateak du giza-puntuazioekin korrelaziorik handiena. HMMetan eta ondorengo probabilitateetan oinarritutako metodo horrez gainera, [45] lanean jatorrizko eta ez-jatorrizko ereduen arteko log-egiantz ratioa (LLR, *log-likelihood ratio*) erabiltzen da ebakera okerreko fonemei antzemateko neurri gisa. Emaitzek agerian uzten dute LLRan oinarritutako metodoak emaitza hobeak dituela ondorengo probabilitateetan oinarritutakoak baino; hala ere, bigarren hizkuntzako hiztunek esandako esaldi zehatzez entrenatu behar da.

1999an, OBEL nazioarteko interesgune bihurtu zen. Urte hartan, teknologia eta hizkuntzen ikaskuntzari buruzko ikerketa eta eztabaida sustatzen zituen *CALICO* aldizkariak ale oso bat eskaini zien teknologia horiei, eta OBELi buruzko lehendabiziko tesi osoa aurkeztu zuen Wittek [46]. Tesi horretan, ebakera-egokitasuna (GOP, textit-Goodness Of Pronunciation) neurria aurkezten da, ondorengo probabilitatearen aldaera bat dena (ikus 6.2. kapitulua). Handik aurrera, GOP neurria oso erabilia izan da ebakeraren ebaluazioan eta ebakera okerreko fonemei antzemateko atazetan.

Wittek proposatutako OBEL sistemaren diseinuaren oinarria 2.2. irudian ageri da. Ahots-sarrerako parametro-erauzketan, ahots-seinalea Mel maiztasuneko koefiziente cepstral (MFCC, *Mel Frequency Cepstral Coefficient*) sekuentzia bihurtzen da, eta MFCC horiek bi ezagutza-pasalditan erabiltzen dira: lehenengo pasaldia lerrokatze behartuko pasaldia da, eta segmentu akustikoen mugak eta trifonemen egiantzak sortzen ditu, Viterbiaren lerrokatzetik kalkulatuak. Bigarren pasaldia fonema-begiztaren pasaldia da, non begizta horretan fonema baten ondoren hurrengoa probabilitate berdinez etor baitaiteke, eta log-egiantzak lehen pasaldian erauzitako segmentazioen gainean lortzen dira. Emaitza horietan oinarrituz, fonemakako GOP puntuazioak kalkulatzen dira. Azkenik, atalase bat ezartzen zaie GOP puntuazioei ebakera okerreko fonemak saihesteko. Atalasea ezartzeko, kontuan izan behar da zer zorroztasun-maila behar den. Atalase egokiaren hautaketa sakonago azaltzen da 6.2. kapituluan.

GOP puntuazioaren zenbait aldaera proposatu dira ordutik. [47]ean, ondorengo log-probabilitate eskalatuaren (SLPP, *scaling log-posterior probability*) metodoa proposatzen da mandarineraz fonema okerrei antzemateko, eta hobekuntza nabaria lortzen da. [48]n, azaltzen da GOPean oinarritutako metodoa eta fonema okerrei antzemateko errore-eskemen detektatzaileak serieko nahiz paraleloko egitura batez konbinatuta,

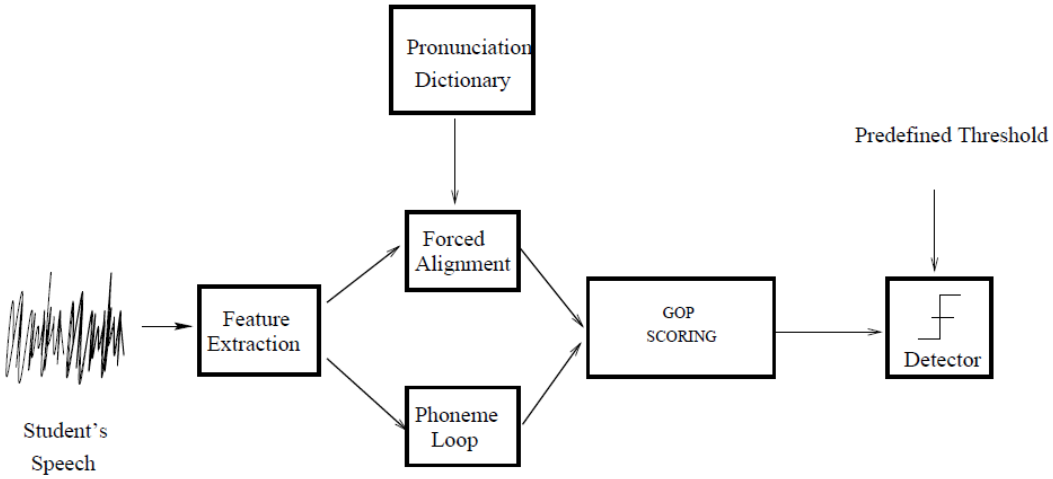


Figure 2.2: Wittek deskribatutako OBEL sistema klasikoaren bloke-diagrama: kontsideratzen da aurrez zehaztutako atalase baten gainetik dauden puntuazioak dituzten fonemek ebakera okerra dutela, eta, beraz, alboratu egiten dira.

batezbesteko errore-tasa txikiagoak eta analisi-*feedback* hobeak lortzen direla serieko egitura erabiliz. GMM-HMMetan oinarritutako ohiko hizketa-ezagutzaileek sortutako puntuazioak hobetzeko, zenbait entrenamendu-algoritmo bereizle ere aplikatu izan dira, besteak beste gehienezko elkar informazioaren estimazioa (MMIE, *Maximum Mutual Information Estimation*) [49], gutxieneko sailkapen-errorea (MCE, *Minimum Classification Error*) [50], gutxieneko fonema-errorea (MPE, *Minimum Phone Error*) eta gutxieneko hitz-errorea (MWE, *Minimum Word Error*) [51]. [52] lanean, MPE bidez hobetutako eredu akustikoak erabiltzen dira ebakera-gaitasuna ebaluatzeko, eta [53] lanean MWE bidez entrenatutako HMM ereduak ere ikertzen dituzte, ingeles ikasleek ebakera okerreko fonemei antzemateko.

Sailkatzaile bidezko planteamenduak ere oso erabiliak izan dira fonema okerrei antzemateko, 2 klasetako sailkapen-ataza bat bezala formulatuta. [54]en, erabaki-zuhaitz batean oinarritutako metodo bat erabiltzen da ebakera okerren mota desberdinetarako atalaseak jartzeko, eta hobekuntza nabarmena lortzen da atalase unibertsal bat erabiltzearekin alderatuz. [55]an ere erabaki-zuhaitzak erabiltzen dira analisi diskriminatzaile linealarekin batera (LDA, *Linear Discriminant Analysis*), nederlanderako *L2* ikasleen ebakera-akats desberdinak hautemateko. Emaitza esperimentalek erakusten dute sailkatzaileetan oinarritutako planteamenduak ondo dabiltzala bokalen ebakera-akatsei antzemateko, baina ez dabiltzala oso ondo kontsonanteen ebakera-akatsekin. Agerian uzten da, orobat, LDAk detekzio-tasa hobeak dituela erabaki-zuhaitzak baino. Lau ikuspegi desberdin konparatzen dira [56]n: GOP puntuazioa, erabaki-zuhaitza, eta LDA bi parametro mota baliatuz: parametro akustiko-fonetikoak eta MFCCak.

Emaitzek erakusten dute LDAn oinarritutako metodoek emaitza hobeak dituztela GOPean eta erabaki-zuhaitzetan oinarritutakoek baino. Hala ere, testa nederlanderako bi fonemekin bakarrik egina dago.

OBEL sistemen funtzionamendua hobetzeko, bektore euskarridun makina (SVM, *Support Vector Machine*) sailkatzaileak ere erabili izan dira, parametro desberdinak konbinatuz: konfiantza neurria, parametro fonetikoak eta MFCCak [57]. [58]n, mandarinerazko silaben ebakera zuzena ala okerra den sailkatzeko, SVM erabiltzen da, ebakeren espazioko ereduak (PSM, *Pronunciation Space Models*) erabiliz ebakera-aldaeren bereizketa hobetzeko. Testuinguru desberdinetan ebakera okerrei antzemateko sailkatzaile gisa SVMa inplementatzen duten beste ikerlan batzuen adibideak dira halaber [59][60][61] eta [62].

Azken urteotan, neurona-sare sakonak (DNN, *Deep Neural Network*) nabarmen hedatu dira. Datuetan maila altuko abstrakzioak modelatzen dituzte, eta nabarmen hobetu dute ASRko eredu akustikoen bereizketa [63]. Egileak, [64]en, azaltzen du DNNetan oinarritutako eredu akustikoetatik estimatutako GOP puntuazioek korrelazio hobea dutela giza adituen ebaluazioekin, GMMetan oinarritutako ohiko sistemetatik erauzitako ohiko GOPek baino. Sistema hobetu bat aurkezten da [65]n, non DNN bidez entrenatutako eredu akustikoak eta, haiekin batera, 2 klaseko erregresio logistikoko sailkatzaileak erabiltzen baitira fonemaka, ingelesezko ebakera-kalitatea neurtzeko. Egileek azaltzen dute sistema horrekin SVMetan oinarritutako puntako sailkatzaileekin baino emaitza hobeak lortzen direla, nahiz eta aldea oso handia ez izan.

Uste-sare sakonak (DBN, *Deep Belief Network*) erabili dira [66]n, ingelesez ebakera okerrak hautemateko eta aztertzeko. Eredu akustikoak modelatzeko, DBN-HMM hizketa-ezagutzeko marko hibridoa baliatu dute, eta hitzen ebakeran errore-tasa erlatibo nabarmen hobea lortu dute, ingelesa ikasteko $L1$en (kantoneraren) mendekoa den corpus batetik abiatuta. Hala ere, GMM-HMMetan oinarritutako sistema klasikoak baino askoz konplexuagoa da konputazionalki.

b) Prosodia-akatsak

Azken urteotan, interes handia piztu da ebakeraren ezaugarri prosodikoak neurtzeko metodo automatikoen ikerketan. Intonazioarekin, azentuarekin eta etorriarekin lotutakoak dira parametro prosodikoak, hala nola f_0 funtsezko harmonikoaren ingerada (malda, batezbestekoa eta maximoa), hitzeko energiaren batezbestekoa, maximoa eta minimoa, silaba azentudun eta azentu gabeen arteko distantziak, hizketa-abiadura, artikulazio-abiadura, fonazio-denbora/-ratioa, fonema-iraupenaren batezbestekoa eta abar.

f_0 ingeradak oso erabiliak izan dira FZa emateko, hala nola *AzAR* sisteman [29] eta *Euronounce* sisteman [34]. 2012an, tesi honetarako lehen pausuak emateko Alemanian egindako ikerketa-egonaldian, f_0 ingerada intonazioaren neurri gisa erabiltzeko baliagarria den aztertzeko ikerketa bat garatu zen [67]. Lan hartan, erreferentziako euskal hiztun baten eta beste 10 hiztunen (6 atzerriko eta 2 euskaldun, ikus 2.1. taula)

$log(f_0)$ kurben arteko Batez Besteko Errore Koadratikoak (BBEK) alderatu ziren. Giza entzumenaren tonu-pertzepzioa maiztasunaren logaritmoarekiko proportzionala denez, maiztasunarekiko berarekiko bainoago, $log(f_0)$ kurbak erabili ziren. Emaitzek erakusten dute $log(f_0)$ kurben artean automatikoki lortutako BBEK distantziak txikiagoak direla euskal hiztunentzat, edozein delarik ere haien generoa (kontuan izan emakumezkoen ahotsak 200 Hz inguruan eta gizonezkoenak 100 Hz inguruan egoten direla). Xehetasun gehiago nahi izanez gero, jo artikulura.

Table 2.1: 8 hiztunen hizketen $log(f_0)$ kurben BBEK, erreferentziazko ahotsarekin konparatuta.

	Ama-hizkuntza	Generoa	Adina	BBEK
01	Japoniera	g	27	0.134
02	Mazedoniera	g	38	0.140
03	Amharera (Etiopia)	g	32	0.134
04	Alemana	g	42	0.165
05	Urdua (India)	g	31	0.129
06	Eslovakiera	e	26	0.135
07	Euskara	e	35	0.111
08	Euskara	g	34	0.113

Etorria ere kontuan hartu beharreko ezaugarria da. [68]en, egileek azaltzen dute ematen duela etorriaren neurrien eta giza gaitasun-epaien artean erlazio lineala dagoela. Orobat, aurkitu da etorriaren giza-puntuazioak fidagarriak direla puntuatzaile artean 0.9tik gorako korrelazioa dutelako [69]. Emaitza horiek agerian uzten dute zer garrantzitsua den etorria neurtzea, ebakera ebaluatzeko edozein ariketaren atal gisa.

Dirudienez, gaur egungo joera da parametro desberdinen multzoak erabiltzea. Adibidez, [70]en, parametro multzo handi bat erabiltzen da hitzen azentuari antzemateko: iraupena, energia, tonua eta isiluneak; lan berriago batean, ikerketa-talde berak erritmo-parametro espezializatuak eta prosodia-parametro orokorrak erabiltzen dituzte prosodia-ebakeraren kalitatea neurtzeko metrika zehatz bat sortzeko [71]. [72]en, egileek etorria lantzeko sistema bat aurkezten dute ikasleen hizketako fonemen iraupenak eta f_0 ingeradak aldatzen dituena, metodo desberdinak erabiliz. Sistemaren asmoa da ikasleari ikusaraztea nolakoa izango litzatekeen bere ebakera. Azterketa pilotu batetik ateratako lehendabiziko emaitzek oso onak dirudite. [73]en, SVMan oinarritutako sailkatzaile bat erabiltzen da tonu-azentua erauzteko.

c) $L1$ekiko mendekotasuna

OBEL ikerketako gai interesgarri bat $L1$ekiko mendekotasuna da. Literaturako ia lan guztiak $L1$en mendekoak dira, tradizionalki $L1$ekiko independenteak diren metodoek baino emaitza hobeak eman baitituzte. Adibidez, *AzAR* sisteman [29] eta *Euronounce*

sisteman [34], hiztun ez-jatorrizkoek egiten dituzten ohiko akatsen zerrenda bat zehazten da, $L1$-$L2$ fonema-nahasmenen pareak sortzeko. [74]en ere, ebakera okerraren arauak eskuz ezartzen dira $L1/L2$ pare jakinetarako, eta erabaki-zuhaitzak erabiltzen dira arauok multzokatzeko.

$L1$en mendeko ikerketa:

Entrenamendurako $L1$ erabiltzeak bi abantaila nagusi ditu: batetik, $L1$ eta $L2$ nahastuz sor daitezke eredu akustikoak [46][75][76]; horrek hobetu egiten du hizketa-ezagutzaren zehaztasuna, eta, beraz, doiago ezagutzen dira ikasleek esandakoak. Hala, malgutasun handiagoa lortzen da ebakera lantzeko ariketak hautatzeko unean. Bigarrenik, ebakera-akatsak berdinak izan ohi dira $L1$ jakin baterako, eta oso desberdinak $L1$ batetik bestera; esate baterako, jatorrizko euskal hiztun batek eta amhareraren edo txineraren jatorrizko hiztun batek ingelesezko ebakera-akats oso desberdinak egingo dituzte. Hortaz, ikaslearen $L1$ zein den jakiteak ebakera-ariketa egokituagoak sortzeko aukera ematen du. Adibidez, [77]en *L1-L2map* izeneko tresna bat aurkezten da; tresna horretan, norvegiera ikasten ari diren $L1$ desberdinetako akatsik ohikoenak daude, eskuz sartuak. Datu horiek ebakera-akatsik ohikoenen zerrenda bat egiteko erabili ziren. Era berean, [78]en, nederlandera-ikasleen akatsik ohikoenen taldeak hautemateko analisi bat aurkezten da, $L1$en arabera.

Metodo desberdinak deskribatu dira $L1$-$L2$ pare jakin bateko akats-eredu ohikoenak hautemateko prozesua automatizatzeko: [79]en eta [80] [81]en, automatikoki sortzen dira ebakera okerren arauak, ebakera kanonikoak eta ez-jatorrizko hizketaren eskuz transkribatutako ebakerak lerrokatuz. Arau horiek ezagutze-sarean inplementatzen dira ebakera-akatsak hautemateko. Horren abantailetariko bat da ezen, akats bat hautematen baldin bada, jakina dela akats mota ere, eta diagnostikoan erabil daiteke datu hori. [82]en, ebakera okerren lexikoak sortzeko bestelako metodo bat aztertzen da: baterako sekuentzien multigramak erabiltzen dituzte 'grafematik ebakera okerrerako' bihurketa egiteko. Azaltzen dute metodo horrekin emaitza pittin bat hobeak lortzen direla, bai zehaztasunari dagokionez, bai alarma faltsuen eta baztertze faltsuen ratioen murrizketari dagokienez. Hala eta guztiz ere, metodo horiek guztiek ere ez-jatorrizko hizketazko corpus eskuz etiketatuak behar dituzte, eta halakoak sortzea garestia eta luzea da.

Witt-ek, 2012an *International Symposium on Automatic Detection of Errors in Pronunciation Training* (IS-ADEPT) kongresuan aurkeztutako ikerketa batean [36], azaltzen du $L1$ekiko independenteak diren metodoei buruzko oso lan gutxi egon direla $L2$ren ezagutzan oinarritutako metodoen antzeko emaitzak dituztenak. Ez-jatorrizko hiztunen datu-baseak grabatzeak eta etiketatzeak oso kostu handia dute, eta gainera ezin dira beste hizkuntza batzuetarako erabili. Horrenbestez, ikasle jakin batentzat haren jatorrizko hizkuntza jakinda akats ohikoenen zerrendak sortzen dituzten metodoak garatzea da erronka, baina haren $L1/L2$ parerako datu-base etiketaturik eduki gabe.

Ondorio gisa, Wittek dio erronkarik handienetariko bat dela $L1$ekiko independenteak diren OBEL sistemak ikertu beharra, edo, gutxienez, $L1$ desberdinetarako erraz konfigura daitezkeenak, eskuz etiketatutako ez-jatorrizko hiztunen datu-baseen beharrik gabe.

Ideia hori ondo egokitzen da tesi honen filosofiarekin; izan ere, xede orokorreko ASRko datu-base bat baino ez dago euskararako OBEL (eta AGP) teknologia garatzeko.

Alemaniako ikerketa-egonaldian egindako lehendabiziko lanean [67], *AzAR* (alemanerako eta hizkuntza eslavieretarako) sistemaren ildo nagusiei jarraiki, euskarazko curriculuma zehaztu genuen, euskara-ikasleek menderatu behar dituzten alde fonetiko eta fonologiko garrantzitsuenak zehaztuz. Hala ere, *AzAR* sistema ez-jatorrizko hizketarekin entrenatzen zen, eta guk ez geneukan aukera hori euskararako. Horren ondorioz, $L1$ekiko independentea den sistema bat garatzeko ikerketari ekin genion.

$L1$ekiko independentziaren ikerketa:

Ebakera puntuatzeko egiantzean oinarritzeak badu abantaila bat: $L1$ekiko independentea da, eta oso erraz kalkulatzen da. Lan askok sailkatzaileak erabiltzen dituzte akats mota ohikoenen fonema pareen kontrasteak modelatzeko (ikus aurreko azpiatalak). Hala ere, badu arazo bat horrek: ezagunak izan behar dute $L1$-$L2$ pare jakin baterako akatsak, eta gainera akats mota bakoitzerako sailkatzaile bat sortu beharra dago.

[67]en, OBEL sistema $L1$ekiko independentea izateko metodo sinple bat aurkeztu genuen. Ebakera okerreko fonemen GOP banaketak sortzeko, sistema-hiztegian aldaketa kontrolatuak txertatzen dira, hau da, posizio jakin bateko fonema bat talde fonetiko beste batekin ordezkatzen da (bokalak, herskariak, sudurkariak, likidoak eta txistukariak), ausaz. Metodo honen oinarrizko ideia da ezen fonema bat oker ebakita dagoela, baldin eta akustikoki hizkuntza bereko beste fonema batetik hurbilago badago. Hala, fonema bakoitzarentzako GOP puntuazioen banaketa-pareak lortzen dira: zuzen ebakitako fonemen banaketa eta oker ebakitako fonemen banaketa. Erabaki-atalaseak lortzeko, bi banaketen arteko errore berdinaren tasa (EER, *Equal Error Rate*) kalkulatzen da (xehetasun gehiagorako, ikus 6. kapitulua). Metodo hau ez da $L1$en mendekoak bezain zehatza, baina abantaila handia da $L1$en mendekoa ez izatea eta $L1$-$L2$ pare jakin bateko akatsik ohikoenei buruzko aurretiko ezagutzarik behar ez izatea.

2.2.2 Ahozko Gramatika Praktika (AGP)

Bigarren hizkuntza ($L2$) eskuratzeko teorien arabera, ez dirudi ikasketa naturalista edo inplizitua nahikoa denik helduek kalitate handiko $L2$ gaitasuna lortzeko; beharrezkoa da ikasketa esplizitua ere [83][84]. Ahozko gaitasunari dagokionez, beste gaitasun batzuetan baino denbora gehiago behar da ikasleari jarduera eta *feedback* nahikoa emateko, zeren banakako irakasle batekin aritu behar baita. Horixe da, hain zuzen, OBHI sistemetan APG inplementatzearen arrazoietariko bat.

Hastapenetako sistemetako estrategia erabilienetako bat da ikasleei esaldi jakin bat esanaraztea, pantailan erakutsitako erantzun multzo itxi batetik aukera bat ozen irakurtzeko eskatuz edo erantzunak sortzeko malgutasun mugatua emanez. Hori modelatzeko, egoera finituko gramatika bat ematen zaio ASRari deskodetzeko, biderik probableena aurkitzeko. Metodo horrek ondo funtzionatzen du, ikasleen akatsak aurrez ikusteko modukoak direnean.

Egoera finituko gramatikak erabiltzearen hastapenetako adibide bat [85]en aurkeztu zuten. Artikuluan, japoniera mintzatua irakasteko helburua duen elkarrizketa-sistema bat deskribatzen da, non hizketa-ezagutza erabiltzen baita elkarrizketako fase bakoitzean ikasleen erantzunak aztertzeko. [86]en, *Subarashii* sistema deskribatzen da, ordenagailuan oinarritutako ahozko hezkuntza-tresna esperimentala. Sistemak arazo sinpleak aurkezten zituen idatziz ingelesez, eta ikasleek ebatzi egin behar zituzten japonieraz esaldi egokia esanez. Egoeren multzoa finkoa zenez, egoera finituko gramatika bat erabiltzen zen, egoera bakoitzerako ahozko esaldi zuzenak eta okerrak ezagutzeko. Sistema hark ASRa erabiltzen zuen ordenagailuaren eta ikaslearen artean komunikazioa bultzatzeko, baina ikasleak ez zuen FZrik jasotzen.

Bilaketa-espazio mugatu batez jardutearen beste adibide bat *Let's go* izeneko OBHI sistema izan zen, 2004an Carnegie Mellon Unibertsitatean garatua ingeleserako [87]. Sistema hark algoritmo bat zuen erabiltzaileak esandakotik ahalik eta hurbilen dauden zuzenketak sortzeko, eta FZa ematen zuen okerreko hitzetan enfasia jarriz. Aurrez, xede-esaldien zerrenda bat ematen zitzaion ikasleari. Gainera, ingelesezko eredu akustikoak egokitu egiten ziren sistemari egindako ez-jatorrizko deien hizketaz, eta hala, estaldura ematen zitzaien Japongo, Indiako, Alemaniako, Txinako eta abarretako hiztunen azentuei ere.

2006ko beste lan batean [88], sorkuntzan oinarritutako bi urratseko markoa erabili zuten: sarrera ez-gramatikal posible baterako, lehen urratsean parafraseatu egiten zen sarrera, eta hitz-sare gainsortu batera egokitzen zen; hala, beharrezko zuzenketak sortzen ziren. Bigarren urratsean, hizkuntza-ereduak eta analisi sintaktikoa erabiltzen ziren berregindako esaldirik onena hautatzeko: hautagai onenen multzo txiki bat sortzen zen, eta, ondoren, testuingururik gabeko gramatika estokastiko bat erabiliz, berriro sailkatzen zen.

Geroago, 2009an, [89]en japonierarako deskribatutako OBHI sistemak ikasleei japoniera ikasten laguntzen zien, ikusizko fitxetan oinarrituta nork bere esaldiak sortuz. Akats lexikoak eta gramatikalak (AGP) hautemateko diseinatuta zegoen, eta ikasleek beren akatsei buruzko FZa jasotzen zuten. Galderak dinamikoki sortzen ziren, ikasgaiko esaldi-moldeen arabera. Erabaki-zuhaitzetan oinarritutako metodo bat ere zuen erantsita, ez-jatorrizko hiztunek egingo zituzten akatsak aurreikusteko, hizketa-ezagutzailerako hitz-gramatikak sortuz.

[90]en, errore sintaktikoak hautemateko metodo berri bat deskribatzen da (nederlanderaz). Ideia nagusia zen ez-jatorrizko hiztunen akats sintaktikoen inbentario bat sortzea, ez-jatorrizko hizketa zuen corpus bateko esaldiak aztertuz. Metodo hark kategoria gramatikalak erabiltzen zituen esaldi bakoitzeko hitzak etiketatzeko, eta algoritmo bat zeukan, bi esalditan hitzak aurkitzeko: xede-esaldian (zuzena) eta esaldiaren errealizazioan (akastuna, beharbada). Informazio hori akatsak hautatzeko eta OBHI sistemetarako ariketak garatzeko erabiltzen zen.

DISCO (Development and Integration of Speech technology into COurseware for language learning) proiektua [91] diseinatu da nederlanderako AGP eta OBEL inplementatzeko eta FZ egoki eta xehea emateko. AGPri dagokionez, ariketa sintaktikoak

ditu inplementatuta, non ikasleak hitz multzo bat ordena sintaktikoki zuzenean esana behar duen. Ariketa horietarako, hizketa-ezagutzaileak erabakitzen du zein den biderik probableena hitz multzoaren permutazio guztiak bide gisa dituen egoera finituko gramatikan zehar. Ariketa morfologikoetan, esaldi osoa aurkezten da pantailan, baina hitz batentzat aukera anitzeko zerrenda bat aurkezten da. [92]en, sistema horretan oinarritutako tresna baten ebaluazioa azaltzen da, non hitz-ordenaren ahozko praktika eskaintzen baita (nederlanderako gramatikarako). Konklusioa da sistema lagungarria dela $L2$n ahozko jarduna lantzeko.

Esakuntzaren egiaztapena (UV, *Utterance Verification*) inplementatu zuten lehen OBHI sistema 2009an aurkeztu zen [93]. Sistema hartan, $L2$ ikasleek nolabaiteko malgutasuna zuten esaldiak esan behar zituzten. Ariketa bakoitzerako aurrez finkatutako erantzun posibleen zerrenda batekin konfiantza-puntuazioak erabiltzen zituen, ASRak erantzun posibleen arteko emaitzarik onena hautatzeko eta, bigarren fase batean, esakuntzaren zuzentasuna egiaztatzeko. Hala, akatsak hauteman zitezkeen, eta akatsok ikasleei erakutsi (FA). Sistemak egiantz-ratioa (ER) erabiltzen zuen konfiantza-puntuazio gisa [94]:

$$LR = \frac{p\left(x|u_1\right)}{p\left(x|u_{FEA}\right)} \tag{2.1}$$

non u_1 deskodetzailearen emaitzarik onena den eta u_{FEA} Fonema Ezagutze Askea (FEA) erabiliz aurkitutako fonema-kate optimoa den, x seinalearentzat. Aurresale horrek adierazten du ezen, sarrerako hizketa bilaketa-espazioko bide batean modelatuta ez dagoenean, $p\left(x|u_1\right)$ egiantza $p\left(x|u_{FEA}\right)$ baino txikiagoa izango dela. Esakuntzaren ondorengo probabilitatea estimatzen du x seinale jakin batentzat, eta $p\left(x|u_{FEA}\right)$ da $x$ren probabilitatearen estimazioa.

Gaur egungo AGP sistemek ez dute denbora errealeko interakziorik eskaintzen. Horrek esan nahi du ikasleak esaldi oso bat esan behar duela FZa jaso aurretik. Hala ere, ikasleak ahoz ariketak ebazten ari den bitartean *feedback*a jasoko balu, zuzenketak berehalakoan egiteko aukera izango luke, esaldi osoa esaten amaitu, zerbitzarira bidali eta *feedback*a jaso beharrik gabe. Horretxegatik aurkeztu da tesi honetan UVean oinarritutako teknika berri bat: Hitzez Hitzeko Esaldi Egiaztapena (HHEE) [95]. Gramatika-ariketak egin ahala ebazteko aukera ematen die ikasleei teknika horrek, zeren, ikasleek, hala, esaldiaren ordenari buruzko beren hipotesia alda baitezakete, zuzena ez den hitz bat esan badu, sistemak une horretan bertan jakinaraziko baitio okerra dela. Xehetasun gehiagorako, ikus 4.3. atala.

2.3 *On-line* inplementazioa

Gure OBHI sistemaren hasierako bertsioa ordenagailu batean instalatu beharreko software bat zen (ikus 6. kapitulua). Horren abantailetariko bat da ez dela Interneteko konexiorik behar; hala ere, softwarea eguneratzeko eta mantentzeko prozesua nekezagoa

da, urrutiko zerbitzari batean instalatutako sistemekin alderatuta. Seinalea zerbitzari batean prozesatzen bada, potentzia handiagoko ordenagailuak erabil daitezke, eta, gainera, prozesaketa-denbora ez da izango prozesadore lokalaren araberakoa. Dirudienez, gaur egungo joera da web (bezero-zerbitzari) konfigurazioa erabiltzea, HTTP zerbitzari batetik web-orri bat jaitsiz eta, audio-datu bitarrak bidaltzeko, *Node.js* zerbitzari batekin konexio bat ezarriz. 7. kapituluan azaltzen da hori.

Web konfigurazioa erabiltzeak kontsiderazio hauek dakartza:

- ***On-line* (denbora errealeko) ahots-aktibitatea detektatzea (VAD, *Voice Activity Detection*)**: Egin ahala ebatzi behar dira HHEE ariketak; hortaz, berehala erabaki behar da zein segmentutan dagoen hizketa eta zeinetan ez. Tesi honetan, metodo berri bat asmatu da: normalizazio anitzeko puntuatzea (MNS, *Multi-Normalisation Scoring*). Metodo horren xedea da hainbat behaketa-egiantz sortzea MFCCak datu-base desberdinetatik kalkulatutako batezbestekoekin eta bariantzekin normalizatuz. Hala lortutako behaketa-egiantzek hizketa-bilbeen eta isilune-bilbeen jokaera karakteriza dezakete baldintza desberdinetan. MNSan oinarritutako VADari buruzko xehetasun gehiago 8. kapituluan aurki daiteke.
- ***On-line* (denbora errealeko) normalizazio cepstrala**: ASRan oinarritutako OBHI sistema bat zerbitzari batean inplementatuz gero, kontuan izan behar da ikasleek zerbitzarira bidalitako audio-seinaleak mikrofono desberdinekin grabatuak izango direla. Hortaz, nolabaiteko normalizazioa aplikatu behar da, sarrerako seinaleen desberdintasun akustikoak konpentsatzeko (kanalak, hondoko zarata eta abar). Ohikoena da erauzitako parametroei batezbesteko- eta bariantza-normalizazio cepstrala (CMVN, *Cepstral Mean and Variance Normalisation*) aplikatzea, baina *on-line* bertsioa behar da audio-bilbeak hizketa-ezagutzaileari batere atzerapenik gabe pasarazteko. MNSan oinarritutako CMVNa proposatu da tesi honetan, emaitza itxaropentsuak dituena. Xehetasun gehiagorako, jo 9. kapitulura.

2.4 Laburpena

Literaturako lan askotan, ikasleek oker ebakitako datuak erabiltzen dira, eredu akustikoak moldatzeko edo zuzen eta oker ebakitako fonemen (edo hitzen) arteko atalaseak kalkulatzeko. Hala ere, OBEL aplikazioak sortzeko berariazko datu-baseak garatzea oso prozesu garestia eta luzea da, eta ikasleen akatsei buruzko aurretiazko lan sakona eskatzen du. Euskara baliabide gutxiko hizkuntzatzat jotzen da, eta gaur egun ez dago OBEL sistemetarako hizketa-teknologiak garatzeko datu-base egokirik. Mikrofonoz grabatutako datu-base akustiko bakarra, ASR oinarri duten euskarazko aplikazioetarako, *Basque Speecon-like* datu-basea da, bulego-ingurunean grabatua eta ikerketarako bakarrik erabilgarria.

Gainera, gaur egungo OBEL sistemen diseinuaren joera, antza, $L1$ekiko independente izatea da. Izan ere, erraza da ikasleen akatsik ohikoenak kudeatzea hizkuntza jakin batean, baina hor jarraitzen du sistema globalago baten beharrak.

AGPari dagokionez, badira ASRan oinarritutako aplikazioak berariaz diseinatuak $L2$ ikasketako zenbait alderdi lantzeko, hala nola morfologia eta sintaxia. Haietariko zenbaitetan, ASR teknologia erabiltzen da ikasleen akats lexiko eta gramatikalak hauteman eta, hala, *feecback*a emateko; beste aplikazio batzuek konfiantza-puntuazioak erabiltzen dituzte, ariketa bakoitzerako aurrez zehaztutako aukera posibleen zerrenda batekin batera. Hala ere, ia ez dago sistemarik ikasleak ariketak ahoz ebazten ari diren bitartean denbora errealeko interakzioa eskaintzen dutenak.

Tesi honetan, ahozko gramatika-ariketak esan ahala ebazteko tresna bat garatu dugu (6. kapitulua), [95]en argitaratua: *Hitzez Hitzeko Esaldi Egiaztapena* (HHEE) izena jarri diogu, eta haren xedea da hizkuntza-ikasle batek esandako esaldi bat hitzez hitz egiaztatzea denbora errealean, egiaztatutako hitza detektatu bezain laster pantailan erakusteko. Kasu horretan, esandako hitza bera da *feedback*a, eta, beraz, garrantzi handia du sistemaren erantzuna zuzena izateak.

Bai OBEL, bai HHEE sistemak urrutiko web zerbitzari batean inplementatu dira, ikusteko ea euskara-ikasleentzat baliagarriak izan daitezkeen ahozko trebetasun eta trebetasun gramatikalak hobetzeko. Horrek zenbait ondorio berri dakartza, hala nola *on-line* (denbora errealeko) VADa, 8. kapituluan azaltzen dena eta [96]en argitaratu dena; eta *on-line* (denbora errealeko) CMVN, 9. kapituluan azaltzen dena.

PART II

Hasierako sistema

CHAPTER 3

Datu-base akustikoa eta fonema-inbentarioa

3.1 Datu-base akustikoa: *Basque Speecon-like* datu-basea

Lan hau aurkezteko unean, *Basque FDB-1060* datu-basea da hizketa ezagutzeko euskarazko sistemak garatzeko eskuragarri dagoen datu-base akustiko bakarra [97]. Datu-base hori Europako *SpeechDat* [98] proiektuaren eskakizunei jarraiki diseinatu zen, eta *ELRA* erakundeak banatzen du bere biltegia[1] baliatuz. *Basque FDB-1060* datu-basea telefonia finkoko sarearen bidez grabatu zen, eta, hortaz, ez da bateragarria, izatez, tesi honen betekizunekin, handik lor daitezkeen eredu akustikoak ez bailirateke optimoak izango mikrofono bidezko audio-sarrera duten sistemetarako.

Hala ere, badira beste datu-base akustiko batzuk euskararako. 2005ean, Eusko Jaurlaritzak *ADITU* izeneko programa jarri zuen abian, hizketa-ezagutzaren eta -sintesiaren euskararako teknologiak garatzeko helburuaz. *ADITU* programaren barnean, bi datu-base berri garatu ziren: *Basque Speecon-like* datu-basea, *Speecon* espezifikazioei jarraituz bulego-ingurunean grabatua, eta *Basque SpeechDat MDB-600* datu-basea, arestian aipatutako *SpeechDat* espezifikazioei jarraituz telefono mugikorraren bidez grabatua. Eusko Jaurlaritzarenak dira biak, eta ez dira publikoki eskuragarriak.

Basque Speecon-like datu-basea bulego-ingurunean grabatu zen; tesi honetako eskakizunekin bat dator, beraz. Datu-basearen espezifikazioak diseinatzeko, *Speecon* proiektua hartu zen abiapuntu gisa, Europako Batzordeak, *Informazio Gizarterako Teknologien* (IST-1999-10003) programaren barnean, Giza Hizketaren Teknologien arloan, diruz lagundutako proiektua. *Speecon* proiektuari 2000ko otsailean ekin zitzaion, hizketa-ezagutzaileak garatzeko baliabideak garatuz. *Speecon*en azken helburua zen merkatuko aplikazioetarako ahots bidezko interfazeak garatzea, eta, hortaz, mikrofono bidez hainbat distantziatara grabatutako audio-fitxategiak behar ziren. Hala, datu-base akustikoak garatu ziren Europako 20 hizkuntzatan, baina euskara ez zegoen zerrenda hartan. *ADITU* programaren helburua horixe izan zen, hain zuzen, euskarazko ASR sistemak garatzeko antzeko datu-base bat sortzea.

1 http://portal.elda.org/en/catalogues/

Datozen ataletan, *Basque Speecon-like* datu-basearen ezaugarri nagusiak azaltzen dira. Xehetasun gehiago nahi izanez gero, ikus [99].

3.1.1 Datu-basearen edukia

Basque Speecon-like datu-baseak bi atal nagusi ditu: irakurritako hizketaz osatutakoa, bata, eta bat-bateko hizketaz osatutakoa, bestea. Irakurritako atala, halaber, bitan zatitu daiteke: hitz gakoak, batetik, eta hitz eta esaldi fonetikoki aberatsak, bestetik. Hona hemen, xeheago:

- **Irakurritako hitz gakoak**: Atal horrek gailu elektronikoari aginduak emateko erabiltzen diren hitzak ditu; adibidez, *aktibatu*, *berrabiarazi*, *aukerak* eta abar. Beste mota batzuetako edukiak ere badira kategoria horretan: digitu-segidak, kantitateak, orduak eta datak, zenbaki naturalak, letreiatzeak, kale- eta hiri-izenen zerrendak, eta zenbait posta-helbide elektroniko eta web-helbide ezagun.
- **Irakurritako esaldi eta hitz fonetikoki aberatsak**: Atal horretan, fonetikoki orekatutako esaldiak eta hitzak daude, egunkarietako, gaur egungo liburuetako eta ahozko transkripzioetako testuez sortutako corpus batetik atereak.
- **Bat-bateko hizketa**: Atal horrek 5 minutu inguruko bat-bateko hizketa du. Batetik, datak, orduak, letreiatzeak, enpresa- eta pertsona-izenak, hiriak, telefono-zenbakiak eta abar ditu, aurrez finkatutako galderen bidez esanaraziak. Bestetik, hizlariaren zaletasunei, filmei eta abarri buruzko kontakizun laburrak ditu, baita telefono bidezko antzezpentxoak ere, banku-transferentziak, hoteletako eta bidaia-agentzietako erreserbak edo zinemako sarreren erosketak antzeratuz.

Datu-basearen edukiak zehazturik ageri dira 3.1. taulan.

Table 3.1: *Basque Speecon-like* datu-basearen edukia (elementu kopurua hizlariko)

	Esaldi/hizlari
Hizketa irakurria	
Xede orokorreko hitz eta esaldiak	32
Xede zehatzeko hitz eta esaldiak	212
Esaldi fonetikoki aberatsak	40
Hitz fonetikoki aberatsak	8
Bat-bateko hizketa	
Bat-bateko elementu askeak	5
Erantzun bideratuak	18

Bat-bateko atalaren eta irakurriaren arteko alde nagusia da bat-batekoan, euskaldun zaharrei dagokienez, aldaera dialektal ugari ageri direla. Hizketa irakurriaren grabazio

ia denak, aldiz, euskara batuan daude. Kontu interesgarria, CALL sistemak garatzeko.

3.1.2 Grabazio-plataforma

Speecon espezifikazioen arabera, ahots-seinaleak lau ingurune desberdinetan grabatu behar dira: bulegoan, etxean, leku publikoetan eta autoan, ASR teknologia garatzean aplikazio-ingurune desberdinak kontuan izateko. Hala ere, Eusko Jaurlaritzak kontsideratu zuen aplikaziorik interesgarrienak bulego-inguruneko baldintzak behar zituztela, eta, hala, gela itxietan egin ziren grabazioak, grabazio-plataforma gisa mahai gaineko ordenagailu bat erabiliz.

Grabazioen konfigurazioa, halaber, sinplifikatu egin zen. Bi mikrofono-kanal baino ez ziren grabatu aldi berean: *hurbilekoa* (entzungailuko mikrofono baten bidez) eta *mahai gainekoa* (hizlariarengandik 1 m-ra jarritako mikrofono baten bidez), nahiz eta *Speecon* estandarrak lau zehaztu: *hurbilekoa*, *papar-mikrofonokoa*, *mahai gainekoa* eta *urrunekoa*. Hurbileko mikrofono gisa*Shure SM10A* bat erabili zen; *mahai gaineko* mikrofono gisa, berriz, *Shure SM58* bat (mikrofono irteeran *Shure FP11* anplifikadore bat jarrita). Audio-seinaleak 16 kHz-etan jaso ziren, 16 biteko PCM kodetzeaz kuantifikatuta.

3.1.3 Datu-basearen tamaina

Datu kopuruari dagokionez, *Basque Speecon-like* datu-baseak 23.8 GB ditu. Dokumentazio-fitxategiek 20 MB dituzte, eta, haien barnean, transkripzio-fitxategiak eta datu-basearen diseinuari buruzko informazio-fitxategiak daude. Gainerako datuak bi bloke berdinetan banaturik daude: batetik, *hurbileko* mikrofonoaren kanalari dagozkion fitxategiak (distantzia laburra); bestetik, *mahai gaineko* mikrofonoaren kanalari dagozkionak (distantzia ertaina). Bloke bakoitzaren tamaina 11.8 GB da.

Datu-baseko fitxategien iraupenari dagokionez, datu-base osoak 109.95 h ditu, hizlariko ia 30 min, batez bestez. Ordu kopuru horretarik, hizketa 52.67 h dira; ez-hizketa (isiluneak, arnas-hotsak eta antzeko soinuak), berriz, 57.28 h. Hizketari dagokionez, 30.37 h hizketa irakurriaren atalari dagozkio, eta gainerako 22.30 h-ak bat-bateko hizketaren atalari.

Bat-bateko hizketaren edukia datu-base osoko hizketaren % 42.34 da. Horrek esan nahi du datu-baseko hizketaren zati handi batek aldaera fonetikoak dituela. 3.2. taulan, datu-baseko atal bakoitzaren ordu kopuruaren laburpena ageri da.

Table 3.2: *Basque Speecon-like* datu-basearen edukia (ordutan)

		h	(h, guztira)
Hizketa	Irakurria	30.37	52.67
	Bat-batekoa	22.30	
Ez-hizketa	Isiltasuna	47.65	57.28
	Arnasa-, mikro-zaratak etab.	9.63	

3.1.4 Hizlarien banaketa, eremu dialektalaren eta hizkuntza-gaitasunaren mailaren arabera

Basque Speecon-like datu-baseko hizlarien banaketa kontu handiz aztertu beharreko kontua da, erabiltzaile potentzialen komunitatearen estaldura behar bezalakoa izan dadin. Euskara oso konplexua da eremu geografikoei dagokienez, zeren eta aldaera fonetiko dialektal ugari ageri baitira, baita euskalki bakoitzaren barnean ere. Euskara batuaren ibilbidea ez da luzea, eta, hortaz, euskaldun zahar asko ez daude ohituta ahozko jardunean euskara batua erabiltzen. Horren ondorioz, euskaldun zahar gehienek nork bere euskalkia erabiltzen dute bat-bateko hizketa grabatzean, hizketa bideratuaren atalean euskara batua erabiltzeko gauza izanik ere.

Eremu dialektalaren araberako eta hizkuntzaren gaitasun-mailaren araberako hizlari-banaketa 3.3. taulan jaso da. Arrazoi historikoak direla-eta, euskaldun berrien kopurua oso handia da Euskal Herrian, eta gaitasun-maila desberdinak dituzte. Datu-basearen dokumentazioan, sailkapen bitar bat ere ageri da hizlarien artean, maila onekoak eta eskasekoak bereiziz. Ikusgai dago datu hori ere 3.3. taulan.

Table 3.3: Hizlarien banaketa, eremu dialektalaren eta hizkuntza-gaitasunaren mailaren arabera *Basque Speecon-like* datu-basean.

	Euskaldun zaharra	Maila ona euskaldun berria	Maila eskasa euskaldun berria	Guztira
Gipuzkoa	85	13	2	100
Bizkaia	49	32	15	96
Nafarroa	14	6	3	23
Araba	0	3	4	7
Gainerakoak	1	2	15	4
Guztira	149	56	25	230

Hizlariak batzeko zailtasunak eta antolaketa-kontuak direla eta, hizlarien azken banaketa ez zen diseinuarenarekin erabat bat etorri. Horren ondorioz, zenbait eremu dialektal (Nafarroa eta Gipuzkoa) diseinuak dioena baino gehiago agertzen dira; beste zenbait, aldiz, gutxiago (batez ere, Araba). Edonola ere, azken banaketa erabilgarria da, eta egokiro irudikatzen du euskal hiztunen demografia.

Aipagarria da Iparraldeko euskalkiak kontuan hartu ez izana *Basque Speecon-like* datu-basearen diseinuan. Hiru euskalki daude Frantziaren eremu administratiboan, Iparraldeko euskalki deritzenak, eta bat ere ez zen kontuan hartu datu-basean. Esan beharrik ez dago ezen, euskararako behar bezalako datu-base akustiko bat egiteko, eremu hartako hiztunen hizketa ere grabatu beharko litzatekeela.

3.1.5 Hizlarien banaketa, adinaren eta sexuaren arabera

Speecon espezifikazioak dio bai helduen bai haurren hizketa jaso behar direla; hala ere, hizlari helduak baino ez ziren grabatu *Basque Speecon-like* datu-baserako. Hizlari kopurua, guztira, 230 da (127 emakumezko + 103 gizonezko). 3.4. taulan, sexuaren banaketa ageri da, estandarrean zehaztutako adin-taldeen arabera banatua.

Table 3.4: Hizlarien banaketa, adinaren eta sexuaren arabera, *Basque Speecon-like* datu-basean.

	Emakumezkoak	Gizonezkoak	Guztira	%
15-30	67	38	105	45,65
31-45	48	51	99	43,04
46+	11	14	25	10,87
Ezezaguna	1	0	1	0,44
Guztira	127	103	230	100
%	55,22	44,78	100	

3.4. taulan ageri denez, *Basque Speecon-like* datu-basean desoreka txiki bat dago emakumezkoen eta gizonezkoen kopuru erlatiboen artean. Hala ere, desbiderapena oso txikia da, eta, hortaz, jo daiteke ez duela ondorio kaltegarririk izango datu-basetik sortutako eredu akustikoetan.

3.1.6 Transkripzio-lanak

Basque Speecon-like datu-basean grabazio guztiak daude; ez, ordea, transkripzio-fitxategi guztiak. Tesi honi ekin aurretik, hainbat lan egin behar izan ziren, hala nola lexikoia egokitu eta bat-bateko hizketa duten grabazioen transkripzioak hobetu.

Transkripzio ortografikok

Hizketa irakurriaren kasuan, audio-fitxategien transkripzio ortografikoak aurrez zeuden eskuragarri. Transkripzio horiek aztertu eta, behar zenean, zuzendu egin ziren. Bat-bateko hizketaren kasuan, transkripzio guztiak sortu ziren eskuz, eta zuzendu egin ziren gero, akatsak konpontzeko eta koherentzia hobetzeko.

Gertaera akustikoak

Transkripzio ortografikoetan, gertaera akustikoak eta hitz-deformazioak ere ageri dira (ikus 3.5. taula). Horrekin, datu-basean ahalik eta hizketa gehienari eusteko aukera ematen da, datu-basetik grabazioak kendu beharra saihesten baita.

Table 3.5: Gertaera akustikoak eta hitz-deformazioak adierazteko erabilitako etiketak, *Basque Speecon-like* datu-basean.

	Ikurra	Esanahia
Ahots-gertaerak	{FIL}	Eten betea
	{FRA}	Hitz zatia
	{LNT}	Hitz luzatua
	{TRC}	Etendako hitza
	{UNI}	Hitz ulertezina
Ahotsa ez diren gertaerak	{BRE}	Arnasa (eta barreak)
	{INT}	Aldizkako zarata
	{SPK}	Hizlariaren soinuak (ezpain-hotsak eta abar.)
	{STA}	Hondoko zarata geldikorra

Transkripzio fonetikoa

Lexikoia hobetu egin behar izan da, sarrera ortografiko guztiekin eta haien transkripzio fonetikoekin. Lexikoi hobetu hori automatikoki sortu da Aholab taldearen euskararako G2P (*grapheme-to-phoneme*) transkribatzailea erabiliz. Gainera, sarrerak euskara batukoak baitira, hainbat ebakera dialektal desberdin ezarri zaizkio hitz bakoitzari, datu-basearen bat-bateko hizketaren atalean ageri diren aldaera dialektalak kontuan izateko.

Hala sortutako lexikoiak 29 626 sarrera lexikal desberdin ditu (euskara batukoak); ebakerak kontuan hartuz, ordez, 122 542 dira. Horrek esan nahi du hitz bakoitzak, batez bestez, 4.14 ebakera desberdin dituela aldaera dialektalak kontuan izateko. Bistakoa da euskararen konplexutasuna, eguneroko bizitzako erabilerari dagokionez.

Basque Speecon-like datu-basean, 36 fonemako oinarrizko multzoa kontsideratu da. Horietariko 35 euskarazko SAMPA alfabetoaren oinarrizko fonema multzoan ageri dira [1] [100]: p, b, t, c, d, k, g, tS, ts, ts‘, gj, jj, f, B, T, D, s, s‘, S, x, G, m, n, J, l, L, r, rr, j, w, i, e, a, o, u. Gainerako fonema ere euskarazko SAMPA alfabetoan ageri da, baina alofono gisa: Z alofonoa, mendebaldeko euskalkietakoa.

3.2 Fonema-inbentarioa

Basque Speecon-like datu-baseko fonema multzoa sinplifikatu egin da lan honetarako. Hasierako 36 fonemetatik, 30 hautatu dira. Atal honetan, azalduko da zergatik hautatu diren fonema horiek eta zein diren haien ezaugarriak. Erabilitako ikur fonetikoak euskarazko SAMPA alfabeto fonetikokoak[2] dira.

1 http://aholab.ehu.eus/sampa_basque.htm
2 http://aholab.ehu.es/sampa_basque.htm

3.2.1 Zenbait kontsiderazio

- **Leherkari ahostunak vs. hurbilkariak**: Fonema leherkari ahostunen eta hurbilkarien ebakera haien kokalekuaren testuinguruaren araberakoa da. *B*, *D* eta *G* hurbilkariak testuinguru-kokaleku jakinetan gertatzen dira, bokal artean, adibidez; *b*, *d* eta *g*, berriz, gainerako kasuetan. Hortaz, fonema beraren alofono kontsidera daitezke. Testuinguruaren araberako fonemak erabiliko direnez, fonema bakarra erabiltzea hautatu da alofono bikote bakoitza adierazteko.
- **Erdi-bokalak**: Fonetikoki bokalen antzekoak diren baina, silabaren nukleoa izan beharrean, silaba-muga diren soinuak dira erdi-bokalak. Diptongoak dira horren erakusgarri garbiak. Euskarazko berezko diptongoak bokal itxiez (*i* eta *u*) amaitzen dira, eta beheranzko diptongoak dira, bokal batez hasi eta erdi-bokal batez amaitzen direnak. Bi erdi-bokal daude, horrenbestez, euskaraz: *j* eta *w*, hurrenez hurren *i* eta *u* bokalei dagozkienak. Inoiz ez dira hiatuak *i* edo *u* bokalez amaitzen diren bokal pareak, bien artean *h* bat izan ezean. Hala ere, *h* fonema, sasoi batean euskararen eremu osoan ahoskatzen bazen ere, Iparraldeko euskalkietan baino ez da esaten gaur egun. Horrek badu ondorioa bat: hiatoa duten hitzak diptongatu egiten dira kasu gehienetan; *ehiza* hitzean, adibidez, transkripzio kanonikoa /*eis‘a*/ bada ere (hiru silaba), maiz /*ejs‘a*/ gisa ahoskatzen da (bi silaba). Diptongazio hori dela eta, oso hiato gutxi aurkitu dira datu-basean, eta, beraz, fonema bakarra (*i* eta *u*) erabili da bokalaz eta hari dagokion erdi-bokalaz osatutako parea adierazteko, zeren testuinguruaren araberakoa baino ez baita.
- *Z* **alofonoa**: Alofono hori, euskarazko fonema-inbentarioko txistukari ahostun bakarra, mendebaldeko euskalkietan agertzen da soilik, oso testuinguru jakinetan: *i* fonemaz amaitzen den hitz bati ’-a’ artikulua ezartzen zaionean, *Z* txertatzen da bien artean. Hori ez da gertatzen mendebaldeko euskalki guzti-guztietan, eta, hortaz, oso lagin gutxi daude datu-basean. Ondorioz, kendu egin da azken fonema-inbentariotik.
- *n* **eta** *l* **fonemen bustidura**: Hegoaldeko euskalkietan, prozesu fonologiko bat gertatzen da euskara baturako ere onartua dena: *n* eta *l* sabaikaritu egiten dira *i* baten ostean kokaturik daudenean. Fenomeno hori beti gertatzen da, zenbait hitz mailegatutan eta izen berezitan izan ezik. Edonola ere, pentsatu da interesgarria dela fonema horien bertsio ez-sabaikarituei nahiz sabaikarituei eutsi eta bakoitzaren eredu akustikoak sortzea; izan ere, lagungarria izan daiteke, aurrez iragar daitekeenez, ebaluazio-lanetan oso ebakera-akats sarria izango denari antzemateko.
- **’*j*’-ren kasua:**: Grafema hori modu desberdinetan ebakitzen da euskaraz, baina euskara batuak *jj* igurzkari ahostuna hobesten du, ekialdeko euskalkietan nagusi dena. Hala eta guztiz ere, oso erabilia da erdialdeko euskalkietako *x* igurzkari belarra, eta, beraz, bi alofonoei eutsi zaie, CAPT atazan erabiltzeko eredu akustiko desberdinak lortzearren.

3.2.2 Azken fonema-inbentarioa

Aurreko guztia kontuan hartuta, tesi honetarako hautatutako azken fonema multzoa 3.6. taulako zerrendan ageri da, ezaugarri nagusien arabera taldekaturik. Ikus euskarazko SAMPA alfabeto fonetikoa, xehetasun gehiago nahi izanez gero[1].

Table 3.6: *Basque Speecon-like* datu-basetik eredu akustikoak sortzeko hautatuko azken fonema-inbentarioa.

Taldea	Fonema
Bokalak	a, e, i, o, u
Leherkari ahoskabeak	c, p, t, k
Albokariak	l, r, rr
Afrikariak	ts‘, ts, tS
Sudurkariak	m, n, J
Sabaikariak	L, jj, gj
Leherkari ahostunak	b, d, g
Igurzkariak	f, x, T, s‘, s, S

3.2.3 Gertaera akustikoen azken zerrenda

Gertaera akustikoei dagokienez, bi hizketa-gertaerari soilik eutsi zaie *Basque Speecon-like* datu-basean: {FIL} eta {UNI} gertaerei. {FRA}, {LNT} edo {TRC} etiketen ordez {UNI} etiketa jarri da, eredu akustikoak sortzeko prozesuan zalantzazko hizketa-segmentu guztiak xurga dezaten. Horrenbestez, {UNI} etiketa ez da geroko ezagutza-esperimentuetan erabiliko, entrenamendu-prozesurako zabor-biltegi gisa baino ez baita erabiliko.

Table 3.7: *Basque Speecon-like* datu-basetik eredu akustikoak sortzeko hautatutako gertaera akustikoen azken zerrenda.

	Ikurra	Esanahia
Ahots-gertaerak	{FIL}	Eten betea
	{UNI}	Hitz ulertezina
Ahotsa ez diren gertaerak	{BRE}	Arnasa (eta barreak)
	{MIC}	Mikrofono-kopeak eta -ukituak
	{SPK}	Hizlariaren soinuak (ezpain-hotsak eta abar)
	{STA}	Hondoko zarata geldikorra

1 http://aholab.ehu.es/sampa_basque.htm

Hizketa ez diren gertakariei dagokienez, lau etiketa erabili dira: *Basque Speecon-like* datu-baseko {BRE}, {SPK} eta {STA} etiketak, eta beste berri bat: {MIC} etiketa, mikrofono-kolpeak, -ukituak eta antzekoak adierazteko balioko duena. Etiketa berri hori sortzearen arrazoia da hurbileko mikrofonoetatik jasotako seinaleak erabiltzea balioetsi dela, tesi honetarako, bai eredu akustikoak sortzeko, bai azken sistema erabiltzeko.

3.7. taulan, hautatutako gertaera akustikoen zerrenda ageri da.

3.3 Konklusioak

Basque Speecon-like datu-basea xede orokorreko euskararako ASR datu-basea da, baina horixe hautatu da tesi honetarako, zeren publikoki eskuragarri dagoen bakarra baita. Datu-baseak bi atal ditu: atal *irakurria* eta *bat-bateko* atala. Nahiz eta atal *irakurria*k zenbait fonemaren ebakeran aldaerak izan, *bat-bateko* atala zeharo dialektala da, eta aldaera fonetiko pilo bat ditu. ASR-rako, aldaera fonetikoak eredu akustiko bakar batez modelatu daitezke. Hala ere, CALL sistema bat egiteko, eredu akustiko "garbiak" behar dira, eta, beraz, hainbat lanketa egin beharra eskatzen du datu-baseak.

Datu-baseak badu abantaila bat: hizlariak hizkuntza-mailaren arabera etiketatuta daude. Hiru etiketa daude hizlariko: jatorrizkoa, goi-maila (ez-jatorrizkoa) eta behe-maila (ez-jatorrizkoa). Informazio hori oso erabilgarria da OBELerako eredu akustikoak sortzeko, jatorrizko hizlariak bakarrik erabiliz, edo OBEL sistema testatzeko, euskara-maila desberdineko hizlariak erabiliz.

CHAPTER 4

Oinarrizko ASR sistema: *AhoSR*

4.1 Sarrera

Kapitulu honetan, OBEL eta AGP sistemen oinarrian dagoen hizketa-ezagutzeko teknologia deskribatzen da: hizketa ezagutzeko *AhoSR* sistema. *AhoSR* hizketa ezagutzeko deskodetzailea da, *Aholab* ikerketa-taldean 2010etik aurrera garatua, eta haren helburua da konputazio-ingurune malgu bat izatea, hizketa-ezagutze automatikoan (ASR, *Automatic Speech Recognition*) oinarritutako aplikazioetarako eta ikerketarako. Funtsean, *C++* lengoaian idatzitako hizketa-ezagutzaile modularra da *AhoSR*. Markoven ezkutuko ereduetan (HMM, *Hidden Markov Models*) oinarriturik dago, eta Mel maiztasuneko koefiziente cepstralak (MFCC, *Mel Frequency Cepstral Coefficients*) erabiltzen ditu parametro akustiko gisa. Hainbat atazatan erabiltzeko diseinaturik dago, hala nola ezagutza fonetikoan, hitz-gramatiketan oinarritutako ezagutzan, eta hiztegi handiko hizketa-ezagutze jarraituan, non hizkuntza-ereduaren informazioa banandurik baitago eta exekuzio-unean eransten baita. Oinarrizko ataza horiei hizketa egiaztatzeko teknikak erants dakizkieke, OBEL eta AGP sistemetan erabiltzeko, batez ere. Deskodetze-prozesua bilbeka garatzen da (*BFS*, *breadth-first search*), lekukoa ematearen paradigmaren bidez [101] eta pasaldi bakarreko izpi-bilaketaren estrategiaren bidez [102]. *AhoSR* plataforma anitzetan erabil daiteke, *Unix* nahiz *Microsoft Windows* sistemetan, eta sarrera gisa, bai zuzeneko audioa (audio-grabagailu baten bidez edo *socket* konexio baten bidez), bai *wav* fitxategiak onartzen ditu. 2014tik, badago*AhoSR*-ren bertsio egonkor bat, [103] artikuluan aurkeztua. Xehetasun gehiago nahi izanez gero, ebaluazio-probak edo aplikazioak, adibidez, ikus artikulua.

Gaur egun, badira kode irekiko zenbait tresna, ASR-ren arloan diharduten ikertzaileentzat erabilgarri; besteak beste, *HTK* [104], *Julius* [105], *Kaldi* [106], *RWTH ASR* [107] eta *Sphinx-4* [108]. Hala ere, badira zenbait gai garrantzitsu geure sistema garatzera bultzatu gintuztenak. Arrazoi nagusia zen aipatutako tresna guztiek dutela moldatu beharra, ASR-ren aplikazio ez-ohiko baterako erabili behar direnean. Adibidez, ASRan oinarritutako OBEL eta AGP aplikazioek egiaztapen-puntuazioak baliatzen dituzte, eta aipatutako tresnetan moldaketa handiak egin beharko lirateke; izan ere, eskuarki,

bilaketa-sare paralelo bat eraiki behar da puntuazio horiek kalkulatzeko (ikus 4.3. atala). Sarritan, aipatutako tresnetan eskua sartu eta moldatzeko ahalegina ezerezetik beste tresna bat sortzearen parekoa izan liteke.

Bestalde, aipatutako tresna horiek hitzean oinarritutako hizkuntz-ereduekin aritzeko optimizatuta daude, eskuarki N-gramekin. N-gramak erabilgarriak dira flexiorik gabeko hizkuntzetarako —adibidez, txinerarako—, flexio arinekoetarako —adibidez, ingeleserako— edo flexio ertainekoetarako —adibidez, espainierarako—. Hala ere, flexio handiko hizkuntzetan edo hizkuntza eranskarietan (euskaran, adibidez, beste askoren artean), hitz-erroei aurrizkiak eta/edo atzizkiak erantsiz eraikitzen dira hitz asko, eta, ondorioz, milioika hitz-forma desberdin sortzen dira, halere maiztasun handikoak direnak [109]. Are gehiago, hain hiztegi-tamaina handiak erabilita, hiztegian ez dauden (OOV, *Out Of Vocabulary*) hitz asko agertzen dira, eta horrek eragin zuzena du ezagutzailearen eraginkortasunean [110]. Hainbat teknika ari dira probatzen gaur egun, batez ere hitza baino unitate txikiagoetan oinarrituak, hala nola turkierarako [111], arabierarako [112], hungarierarako [113], tamilerarako [114] eta euskararako [115]. Hitza baino unitate txikiagoetan oinarritutako hizkuntza-ereduak erabilita, moldatu egin behar da bilaketa-espazioa, propagazio-bideak doi-doi kontrolatuko badira.

Erabilera komertzialerako aukera ere kontuan izan behar da, desberdina baita tresna batetik bestera. ASRan oinarritutako OBEL edo AGP tresna bat, adibidez, dagoeneko egina dagoen hizkuntzak ikasteko tresna batean integratu behar bada, edo sistema kapsulatu batera migratu beharra ikusten bada, ezinbestekoa da kode osoa erabiltzeko aukera izatea. Hori guztia kontuan izanda, kontsideratu genuen egokia litzatekeela ezagutze-sistema moldagarri bat garatzea, teknika desberdinak aplikatu eta testatu ahal izateko; gainera, hala errazagoa litzateke etorkizuneko garapenekin ere jardutea.

Kapitulu honetan, lehendabizi *AhoSR*-ren sistema-arkitektura orokorra aurkezten da, eta bloke bakoitza xehetasunez deskribatzen da. Ondoren, Hitzez Hitzeko Esaldi Egiaztapenaren (HHEE) atazarako egin behar izan diren aldaketak azaltzen dira: egiaztapenerako grafo paralelo bat eranstea, esaldietarako bilaketa-grafoa hobetzea eta moldatzea, bai eta bat-bateko egiaztapen-prozesuan erabakiak hartzeko prozedura.

4.2 Sistema-arkitektura

AhoSR-ren arkitektura, oro har, modularra da; hala, errazago egin daitezke moldaketak eta egokitzapenak bloke bakoitzean, gainerakoetan eraginik izan gabe. Lau modulu nagusi daude *AhoSR*n: *Kudeatzaile Nagusia*, *Audio-sarrera*, *Ezagutza Linguistikoa* eta *Deskodetzailea*. *Kudeatzaile Nagusia*k ezagutze-prozesuaren parametroak ezartzen eta kudeatzen ditu, eta ezagutzailearen atal desberdinen exekuzio-sekuentzia kontrolatzen du. Halaber, *Deskodetzaile*tik jasotako datuak prozesatzen ditu, azken emaitza sortzeko. *Audio-sarrera*k parametrizatu egiten du sarrerako audio-seinalea (bai zuzenekoa, bai *wav* bidezkoa) eta sortutako parametro bektoreak *Deskodetzaile*ra bidaltzen ditu. *Ezagutza Linguistikoa*ren atalean, hiru datu mota gordetzen dira: eredu akustikoak, lexikoia eta hizkuntza-eredua, zeinak bilaketa-grafoa eraikitzeko erabiltzen baitira. *Deskodetzailea*k, lehenik, atazarako bilaketa-grafoa sortzen du; jarraian, parametro bektoreak hartzen ditu

*Audio-sarrera*tik, eta deskodetze-prozesua abiarazten eta kontrolatzen du bilaketa-grafo horretan zehar. Deskodetzeari buruzko xehetasunak eta emaitzak *Kudeatzaile Nagusia*ri bidaltzen dizkio. *AhoSR*-ren bloke-diagrama orokorra 4.1. irudian ageri da.

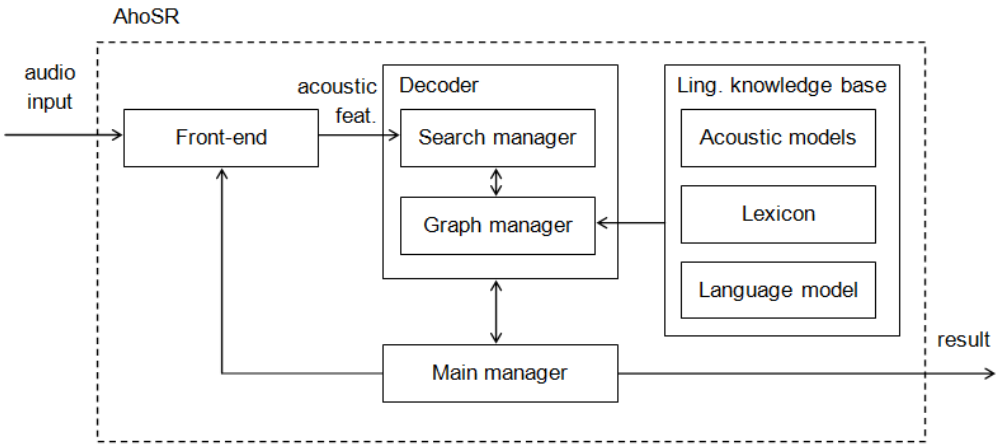


Figure 4.1: *AhoSR*-ren sistema-arkitektura. Bloke nagusiak hauek dira: *Kudeatzaile Nagusia*, *Audio-sarrera*, *Ezagutza Linguistikoa* eta *Deskodetzailea*. Irudian, blokeen arteko komunikazioa ageri da.

4.2.1 Kudeatzaile Nagusia

*Kudeatzaile Nagusia*ren ardura nagusia da atal desberdinen arteko komunikazioa kudeatzea. Lehendabizi, konfigurazio-fitxategi bat irakurtzen du, erabiltzaileak hainbat parametroren balioak finkatzeko erabiltzen dena, eta balio horien artean bateraezintasunik baden aztertzen du. Balio horrek kontuan hartua, *Kudeatzaile Nagusia*k modulu bakoitza konfiguratzen eta abiarazten du, eta, orobat, erabakitzen du deskodetze-prozesuak zer exekuzio-sekuentzia jarraitu behar duen. Orduan, deskodetze-prozesua abiarazten du.

Deskodetze-prozesuan, *Kudeatzaile Nagusia*k ezagutzailearen atalen arteko komunikazioa kudeatzen du, eta, deskodetze-prozesua amaitu denean, *Deskodetzaile*tik jasotako datuak prozesatzen ditu eta azken emaitza prestatzen du. Emaitzak pantailan erakutsi edo gorde egin daitezke, erabiltzaileak hautatutako formatu batean: iraupenekin edo gabe, probabilitateekin edo gabe, fonema-mailan edo hitz-mailan, egiaztapen-puntuazioekin edo gabe, eta abar.

Konfigura daitezkeen parametrorik garrantzitsuenak A. eranskinean ageri dira.

4.2.2 Audio-sarrera

*Audio-sarrera*ren helburu nagusia da parametroak erauzteko prozesua kudeatzea. Prozesu horretan, sarrerako audio datuak MFCC bektore bihurtzen dira, eta delta eta delta-delta koefizienteak (lehen deribatu eta bigarren deribatuak) ere erants daitezke, dinamikaren informazioa gal ez dadin [116]. Hainbat parametroren balioak ezar daitezke, MFCC desberdinak lortzeko: bilbe-tasa, bilbe-luzera, *mel* edukiontzien kopurua, goiko eta beheko

ebakitze-maiztasunak eta abar. Koefizienteak sortzeko prozesua eta haiek gordetzeko pilaren kudeaketa *Kudeatzaile Nagusiak* kontrolatzen ditu, zeinak *Deskodetzaile*ari mezu bat bidaltzen baitio datu berriak erabilgarri daudenean.

Gainera, *Audio-sarrera*k beste teknika lagungarri batzuk ere baditu inplementatuta, hala nola ahots-aktibitatea detektatzea (VAD, *Voice Activity Detection*) (ikus 8. kapitulua) eta batezbesteko eta bariantza normalizazio cepstrala (CMVN, textitCepstral Mean and Variance Normalisation) zaratarekiko funtzionamendu sendoagoa izan dezan (ikus 9. kapitulua).

4.2.3 Ezagutza Linguistikoa

*AhoSR*n, hiru ezagutza-iturri erabiltzen dira, ezagutzailearen deskodetze-prozesuko bilaketa-grafoa sortzeko:

Eredu akustikoak

HMMak [117] erabiltzen dira, hizketa-seinalearen egitura akustikoaren sekuentzialtasuna modelatzeko; unitate bakoitzaren aldakortasun espektrala Gauss-en dentsitate-nahasteak erabiliz modelatzen da (dentsitate jarraituko HMMak).

HMMek bi parametro mota dituzte: trantsizio-probabilitateak eta igorpen-probabilitateak (edo irteerako probabilitateak). Trantsizio-probabilitateek kontrolatzen dute nola aukeratzen den j egoera t unean, k egoeratik abiatuta $t-1$ unean: $a_{kj} = p(s_j|s_k)$. Igorpen-probabilitatea, berriz, s_j egoeran egonik x obserbazioa sortzeko probabilitatea da: $b_j(x) = p(x|s_j)$, eta sarrerako hizketa-bektorearen parametro bakoitzarentzat hari dagokion nahaste-elementu bakoitzaren probabilitatea kalkulatuz lortzen da. 4.2. irudian, hiru igorpen-egoerako HMM generiko baten topologia ageri da.

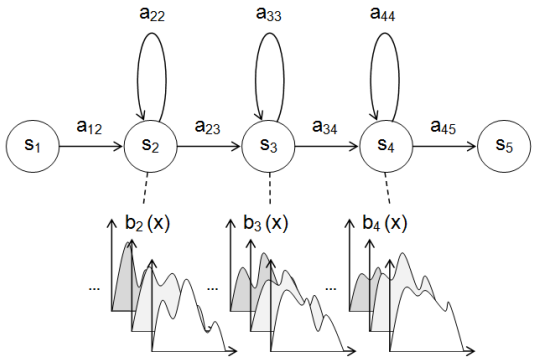


Figure 4.2: *AhoSR*k dentsitate jarraituko HMMak baliatzen ditu eredu akustiko gisa. Irudian, hiru igorpen-egoerako HMM baten topologia ageri da, a_{ij} trantsizio-probabilitateak eta $b_j(x)$ irteerako pdf-ak dituena.

AhoSR hitzetan nahiz azpi-hitzetan oinarritutako HMMak erabil ditzake. Hala ere, trifonemak erabiltzeko optimizatuta dago, fonemen testuingurua kontuan hartzen baitute. Trifonema HMMen arazoetako bat da oso eredu kopurua handia sortzen dela erabilgarri izan ohi diren entrenamenduko datuetarako. Gainera, trifonemen testuinguru asko

oso antzekoak dira. Horrenbestez, *AhoSR*k aukera ematen du egoera batuko HMMak erabiltzeko. Egoera batuei *senone* ere deritze, eta antzeko egoerak batuz sortzen dira [118]. Hala, egoera bakoitzaren irteera-banaketako parametroak hobeto estimatzen dira entrenamenduko materialaz, eta, gainera, nabarmen murrizten da prozesaketa-denbora.

Eredu akustikoen formatua HTK tresnarekin bateragarria da. Horrek esan nahi du HTK-rekin sortutako HMMak *AhoSR*n ere erabil daitezkeela.

Lexikoia

Hitzaren ebakeraren eta idatzizko adierazpenaren arteko erlazioa duen fitxategia da lexikoia. Ebakera adierazteko, unitate segidak (hitzak, silabak, fonemak eta abar) erabiltzen dira. Unitate bakoitza HMM bati dagokionez, lexikoiko hitz bakoitza HMM egoeren segida batez adierazirik geratzen da (ikus 4.3. irudia).

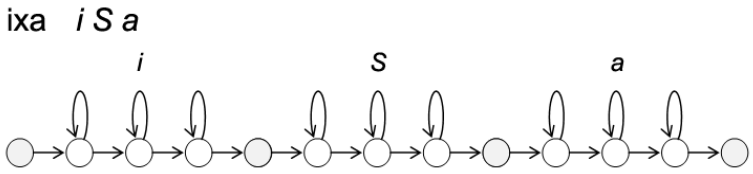


Figure 4.3: Euskarazko **ixa** hitza, lexikoian, /*i S a*/ HMM segida gisa adierazten da (euskararako SAMPA kodea).

*AhoSR*k, gainera, hitzen ebakera anizkoitzak maneiatzen ditu; hala, aldaera dialektaletako eta bestelako ebakera-aldaerak ere har daitezke kontuan. Lexikoian, sarrera desberdin gisa ezarri behar dira, hitz lexiko berarekin (eta dagokion ebakeraz).

Hizkuntza Eredua

Bi hizkuntza-eredu mota maneiatzen dira: testuingururik gabeko gramatikak eta *N*-grametan oinarritutako hizkuntza-ereduak. *AhoSR*n testuingururik gabeko gramatiketarako hautatu den estandarra BNF (Backus-Naur Form) areagotuko notazioa da [119], zeinak hizketa-ezagutzarako gramatiken sintaxia zehazten baitu. Bestalde, hizkuntza-eredu estatistikoa inplementatzeko, ARPA formatuko *N*-gramak ere erabil daitezke [120], adibidez SRILM tresnarekin sor daitezkeenak [121].

AhoSR-ren izaera modularraren ondorioz, sisteman gramatika-formatu berriak erants daitezke bilaketa-espazioaren barne-errepresentazioa sakonki ezagutu beharrik gabe. Horrek aukera ematen du beste ataza batzuetarako ebazpide berriak testatzeko edo hizkuntza eranskariak (euskara, adibidez) ikertzeko.

4.2.4 Deskodetzailea

Deskodetzailea bi modulu nagusiz osatuta dago: *Grafo-kudeatzailea*, behar den atazarako bilaketa-grafo egokia nola eraiki kontrolatzen duena; eta *Bilaketa-kudeatzailea*, *Audio-sarrera*tik parametro akustikoak jaso eta *Grafo-kudeatzailea*k sortutako grafoan zeharreko deskodetze-prozesua kudeatzean duena. *Bilaketa-kudeatzailea*k lortutako emaitzak *Kudeatzaile Nagusia*ri bidaltzen zaizkio.

Grafo-kudeatzailea

Haren funtzio nagusia da atazarako bilaketa-grafo egokia sortzea. Lehendabizi, *Ezagutza Linguistiko*ko informazioa barneko data-egitura batera itzultzen du. Ondoren, bilaketa-grafoa sortzen du *Ezagutza Linguistiko*ko informazioaren bidez: hizkuntza-ereduko informazioa erabiliz, *Grafo-kudeatzailea*k hitz-mailako sarea sortzen du, nodoz eta arkuz osatua. Nodoek hitzak adierazten dituzte eta arkuek, berriz, nodoen arteko erlazioak. Jarraian, sare horretako nodo bakoitza lexikoiak dakarren HMM adierazpenen sekuentzia egokiaz ordezkatzen da (kontuan hartzen dira ebakera anizkoitzak). Azkenik, HMM adierazpen bakoitza hari dagokion HMMarekin lotzen da; hala, egoera-mailako sarea edo azken bilaketa-grafoa lortzen da, nodoz eta arkuz osatua. Ezagutze fonetikorako, bilaketa-grafoa sortzeko modu berezi bat erabiltzen da, HMMek hitzak adieraziko balituzte bezala kontsideratuz.

Bilaketa-grafoa konprimituz, nabarmen jaits daiteke bilaketa akustikoaren esfortzua. *AhoSR*k aukera ematen du bilaketa-espazioan aurretiko nahiz atzetiko zuhaitz-konpresioa ezartzeko [122][123]. Bilaketa-espazioaren topologia horretan, Hizkuntza-ereduarena modulu bereizia da, *Bilaketa-kudeatzailea*k exekuzio-denboran kontsultatzen duena. Ezaugarri hori dela eta, memoriarekiko eraginkorra ez ezik, erabileran malgua ere bada *AhoSR*.

Hizketa-egiaztapena ere behar baldin bada, *Grafo-kudeatzailea*aren ardura da bilaketa-grafo paralelo bat sortzea, bertan egiaztapen-puntuazioak kalkulatzeko.

Bilaketa-kudeatzailea

*Bilaketa-kudeatzailea*k lekukoa ematearen algoritmoa darabil [101] deskodetze-prozesurako. Lekukoak bilaketa-grafoan zehar hedatzen dira, Viterbi algoritmo estandarra baliatuz. Tokenek bilaketari buruzko informazioa gordetzen dute, bilaketako bide aktibo guztien historia osatuz. Gainera, lekuko bakoitzak bideko puntu jakin bateko puntuazio akustiko eta linguistiko orokorrak gordetzen ditu. Bilaketa aurrera joan ahala, sarrerako parametro-bilbe bakoitzari puntuazio bat ezartzen zaio, lekuko-egoera bakoitzari lotutako eredu akustikoak baliatuz, eta inausi egiten dira puntuazio baxuko adarrak. Bi mota inausketa mota inplementatu dira (eta konbinatu egin daitezke): tarte-inausketa orokorra, zeinak bide-hipotesi partzial onenetik hurbil dauden egiantz-puntuazioei soilik eusten baitie; eta histograma-inausketa, zeinak mugatu egiten baitu denbora-bilbe bakoitzeko bide aktibo kopurua, bide onenen kopuru finko bati bakarrik eutsiz [124]. Konfiguratu egin daiteke grafoko nodo-egoera bakoitzean edo nodo-lagungarri bakoitzean zehar hedatzen den lekuko kopurua; hala, $N - Best$ zerrenda eskuratu daiteke, adibidez.

Egiaztapeneko funtzionamendu-modua hautatuz gero, *Bilaketa-kudeatzailea*k fonema baten edo fonema-sekuentzia baten ebakera-egokitasun (GOP, *Goodness Of Pronunciation*) puntuazioak ere kalkulatzen ditu, denboraz normalizatutako ondorengo probabilitatearen logaritmo gisa segmentu akustiko batean zehar [125]. GOPa kalkulatzeko, bi puntuazio sorta erabiltzen dira: batetik, ezagutze-prozesuak dirauen bitartean bilaketa-grafo nagusian lortzen direnak; bestetik, fonema askeen begizta batez osatutako grafo sekundario batean lortzen direnak (ikus 4.3. atala).

4.3 HHEE atazarako egokitzapenak

Sarreran azaldu denez, *AhoSR* egokitu egin behar izan da, ibili ahalako HHEE ataza inplementatzeko. Lehenik, grafo paralelo bat erantsi behar izan da, egiaztapen-puntuazioak kalkulatzeko. Bigarrenik, hobetu egin da esaldiak egiaztatzeko bilaketa-grafoa, hitzen arteko koartikulazioa eta hitz batek izan ditzakeen ebakera desberdinak kontuan hartuz. Azkenik, moldatu egin da *Bilaketa-kudeatzailea*, egiaztatzeko ea hitz batek noiz gainditzen duen atalase-balio bat, eta hala erabakitzeko ea hitza egiaztatutzat jotzen den ala ez. Egiaztatutzat joz gero, prozesu bera abiarazten da esaldiko hurrengo hitzean. Atal honetan, xeheago azaltzen da atal bakoitza zertan den.

4.3.1 Grafo paraleloa

6.2. atalean azalduko denez, oinarrizko GOP neurria (4.1) ekuazioa erabiliz kalkulatzen da.

$$GOP\left(q_i\right) \approx \frac{1}{T_i} log\left[\frac{p\left(O_i|q_i\right)}{p\left(O_i|q_{j_{max}}\right)}\right] \tag{4.1}$$

non j_{max} ebaluatu beharreko segmentuan egiantz altuena ematen duen fonema-ereduaren indizea baita. Hortik abiatuta, bilaketa-grafo nagusiarekiko paralelo korritzen den begizta aske bat erabiliz kalkulatzen da GOP balioa. Begizta aske hori osatzeko, deskodetze-prozesuko trifonema-eredu guztiak hartu beharko lirateke kontuan, baina horrek, trifonema-eredu kopurua oso handia baita, atzerapen nabarmena eragingo luke sisteman. Horren ordez, monofonema-begizta bat erabili ohi da, 30 eredu besterik ez baitira. Hala, begizta paraleloa arina da eta ez du inolako atzerapenik eragiten.

4.3.2 Bilaketa-grafo berezia

HHEE atazan, erabiltzaileak esaldi bat pentsatu eta eraiki behar du denbora-tarte jakin batean. Hori horrela, deskodetze-prozesuko bilaketa-grafoak hiru egoera desberdin kontsideratu behar ditu hitz batetik hurrengorako jauzian.

- **Isilunea hitzen artean**: Erabiltzaileal geldialdi bat egiten du, eta isilune bat txertatzen da hitzen artean.
- **Ondoz ondoko hitz koartikulaziogabeak**: Erabiltzaileak eten labur-laburra egiten du hitzen artean, isilunetzat jotzeko baino laburragoa, eta koartikulazioa ere eten egiten da.
- **Ondoz ondoko hitz koartikulaziodunak**: Ondoz ondoko hitzak inolako etenik gabe ebakitzen dira, eta koartikulazioa gertatzen da hitzon artean.

4.4. irudian, hiru egoera horiek kudeatzeko diseinatutako bilaketa-grafo baten adibide bat dago. Irudian, hiru hitzeko esaldi bat ageri da ("asteartea, osteguna, larunbata"), eta kontuan hartzen ditu hitz bakoitzaren ebakera desberdinak.

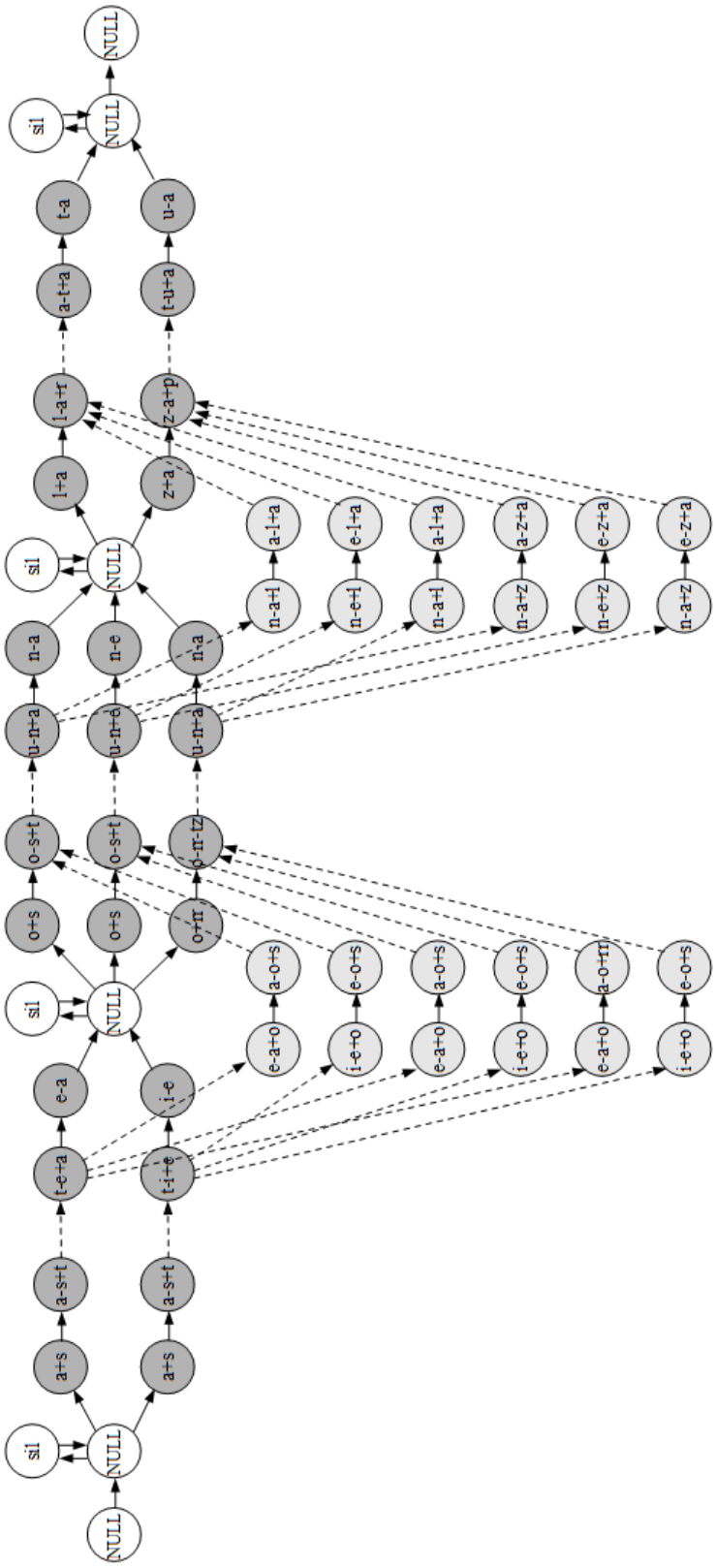


Figure 4.4: "Asteartea, osteguna, ostirala" esaldiarentzako deskodetze-sare baten adibidea.

Ingurune errealista batean, ikasleak akatsak egingo ditu. Horrek esan nahi du egiaztapen-sistemak halako ahots-segmentuak maneiatu behar dituela, haiek Viterbi deskodetzailean duten eragina xurgatuko bada. Sistemak, kasu horretan, espero baino ahots-bilbe gehiago jasoko ditu, eta hori modelatu egin behar da nolabait. Hala, *AhoSR*-ren bilaketa-grafo nagusiari fonema-begizta lagungarriak erantsi zaizkio hasieran, amaieran eta hitzen artean, 4.5. irudian adierazi bezala. Erabiltzaileak egindako akatsak baleude, fonema-begizta horiek gehitu ezean okerra izango litzateke Viterbi algoritmoak emandako segmentazioa, eta egiaztapena edo puntuatzea ez litzateke segmentazio hori oinarri hartuta kalkulatuko.

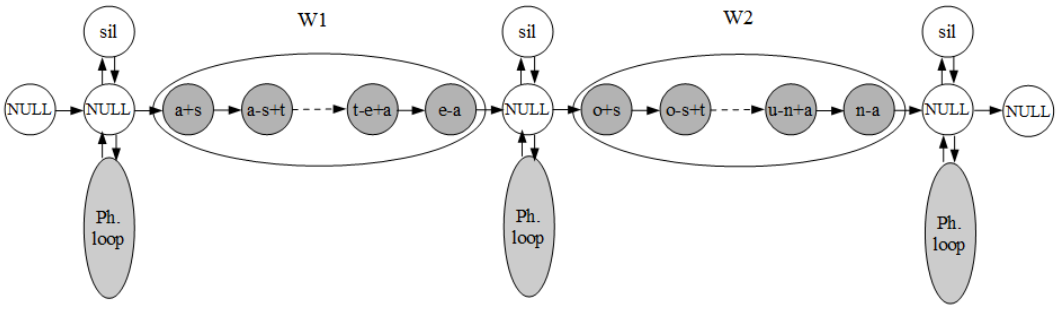


Figure 4.5: HHEE atazarako, *AhoSR*-ren bilaketa-grafoko isilune-nodoei paraleloan erantsitako fonema-begizta askeak.

4.3.3 *Bilaketa-kudeatzaile*ko egokitzapenak

*Bilaketa-kudeatzailea*k erabaki behar du egiaztatzen ari den hitza benetan hitz hori dela. Hori egiteko, sistemak hitz mailan ebaluatu behar ditu konfiantza-puntuazioak. Aurrez finkatutako atalase bat gainditzen duen puntuazio altuago bati antzematen bazaio konfiantza-puntuazioen kurban, egiaztapen sakonagoa egiten da, fonema-mailakoa alegia, Viterbi deskodetzaileak emandako segmentazioa baliatuz. Fonema bakoitzak berari dagokion atalasea gainditzen baldin badu, egiaztatutzat jotzen da hitza, eta berehala erakusten da pantailan. Orduan, eguneratu egiten dira egiaztapen-prozesuarekin lotutako bilaketa-grafoko baliabideak, eta hurrengo hitza prozesatzeari ekiten zaio.

Fonema batek ere ez balu atalasea gainditukо, uneko bilbea baztertu egiten da, eta egiaztapen-prozesuak hurrengo bilbearekin jarraitzen du. HHEE prozesua amaitutzat jotzen da azken hitza positiboki egiaztatzen denean.

4.3.4 Audio-sarrerarako *socket*a

Audio-sarrera *wav* fitxategi bat izan daiteke, edo, orobat, audioa grabatzeko gailu batetik edo *socket* konexio batetik datorren zuzeneko audio-fluxua. Aipatzekoa da tesi honetarako berariaz erantsi dela *socket* bidezko sarreraren funtzionaltasuna, *AhoSR* Internet bidez atzitu eta denbora errealeko funtzionamendua inplementatzeko. Funtzionaltasun horrek

aukera eman digu *AhoSR* zerbitzari batean instalatzeko, eta, ondorioz, modu publikoan atzi daitezkeen zenbait *demo* garatu ditugu, ez bakarrik HHEEn oinarritutako AGP atazetarako, baizik eta ezagutze-atazetarako ere bai. Gune honetan probatu daitezke: http://aholab.ehu.eus/users/igor/demos.html.

4.4 Konklusioak

Kapitulu honetan, *AhoSR*-ren egitura eta funtzionaltasunak deskribatzen dira. Tesi honetan erabili den hizketa-ezagutzako oinarrizko softwarea da *AhoSR*, *Aholab* ikerketa-taldean sortua eta garatua. HMMetan oinarrituta dago, eta MFCCak darabiltza parametro akustiko gisa. Hainbat ataza kudeatzeko diseinaturik dago, hala nola ezagutza fonetikoa, hitz-gramatiketan oinarritutako ezagutza eta hiztegi handiko hizketa-ezagutze jarraitua.

Tesi honen xedeetarako, hizketa egiaztatzeko teknikak inplementatu dira, oinarrizko atazekin batera exekutatzeko. Horretarako, grafo paralelo bat inplementatu da, GOP puntuazioak kalkulatzeko. OBEL eta HHEE atazetarako ere berariazko bilaketa-grafoak erantsi dira, bakoitzaren beharrizanetara egokitzeko.

Gainera, *socket*ak inplementatu dira *AhoSR*-ren audio-sarrerako moduluan. Horrek denbora errealean funtzionatzeko aukera ematen du ezagutzailea Internet bidez atzitzen denean, eta, beraz, *AhoSR* zerbitzari batean instalatzeko aukera ematen digu horrek, sarbide unibertsala bermatuz.

AhoSR-ri buruzko xehetasun guztiak [103]en aurki daitezke.

CHAPTER 5

Eredu akustikoak: HMMak

5.1 Sarrera

Eredu akustikoak funtsa dira, bai Hitzez Hitzeko Esaldi Egiaztapena (HHEE) atazarako, bai ebakera ebaluatzeko atazarako. GOP puntuazioak lortzeko, biek erabiltzen dituzte bilaketa-grafo nagusi bat eta egiaztapen-grafo lagungarri bat, Markoven ezkutuko ereduak (HMM, *Hidden Markov Models*) uztartuz eraikiak. Hortaz, azken emaitzetarako, erabakigarria izango dira HMMak (funtsean, behaketa-egiantzez eta trantsizio-probabilitatez osatuak) eta haien kalitatea.

Kalitatezko eredu akustikoak lortuko badira, behar bezala etiketatutako corpusa behar da. 3.1. atalean azaltzen denez, euskararako dagoen mikrofono bidez grabatutako datu-base akustiko publiko bakarra *Basque Speecon-like* datu-basea da [99]. Datu-base horrek 230 saio ditu, bat hizlari bakoitzeko, eta bi bloketan banatu da, lan honetako esperimentuak egiteko: *train* edo entrenamendu-blokea, lehen 155 saioez osatua (74.10 h); eta *test* edo saiakuntza-blokea, gainerako 75 saioez osatua (35.85 h). *Train* blokeaz entrenatu HMMak, eta handik erauzi dira 6. ataleko GOP puntuazioak; *test* blokeaz ebaluatu dira.

Datu-basearen *Bat-bateko hizketa*ren atala hizketa dialektalaz osaturik dagoenez eta hizketa-eduki osoaren % 42.34 denez (ikus 3.2. taula), ebakera alternatiboak sortu dira, aldaera horiei aurre egiteko. Ebakera alternatiboak oso erabiliak dira hizketa ezagutzeko esperimentuetan, eta eskuarki eskuz edo egiantz handieneko ikaskuntzaz lortzen dira. Alde batetik, handiagotu egiten dute ebakera-aldakortasunaren estaldura; bestetik, areagotu egiten dute elementu lexikalen arteko nahasmena. Kontrako bi faktore horien ondorioz, eskuarki ez da ezagutzan funtzionamendu-hobekuntzarik lortzen, edo oso txikia, hiztegi estandarrak erabiltzearekin konparatuz (adibidez, [126]). Hala ere, lan honen helburua ez da eredu onak sortzea ezagutzarako, baizik eta ebakera ebaluatzeko. Hortaz, planteamendu hori egokia izan daiteke, edo, behinik behin, hasierako ikerketa-lana izan daiteke, kalitate oneko eredu akustiko bereizleak entrenatzeko.

Eredu akustikoen kalitatea neurtzeko, fonemen errore-tasa (PER, *Phonetic Error Rate*) erabiliko da. Gure helburua izango da gaur egun beste hizkuntza batzuetarako

argitaratu diren PER balioetatik ahalik eta hurbilen egotea. Ezagutze fonetikoari buruzko atzera begirako errepaso bat eginez gero, 2011n argitaratutako artikulu batean deskribatzen da zer emaitza lortu diren 1990etik 2010era bitartean TIMIT datu-basea erabiliz [127]. Han, azaltzen da ezagutze fonetikoaren emaitzak % 13 inguru hobetu zirela 1990 eta 2010 bitartean, eta hori lehen 5 urteetan zehar gertatu zela batez ere: 1990ean % 26.20ko gutxieneko PERa lortu zen, monofonema HMM diskretuak erabiliz [128], eta 1995 inguruan PERaren gutxieneko balioa % 22.50 ingurura jaitsi zen, trifonema HMM jarraituak [129] eta sare errepikakorrak [130] erabiliz. Handik 2011ra bitartean, oso hobekuntza txikiak lortu ziren. Azterketa hartako konklusioek zioten zaila zirudiela % 20ko muga gainditzeak; garai hartako ikuspuntuaren isla zen hura. Hala ere, urte hartan bertan, laster ikusi zen hobekuntzak etorriko zirela, neurona-sare artifizialei esker (ANN, *Artificial Neural Network*). 2015ean % 17.10eko PERa lortu zutela argitaratu zuten [131]-n, neurona-sare konboluzionalak (CNN, *Convolutional Neural Networks*) erabiliz, eta, urte hartan, beranduago, autore berak % 16.5eko PERa lortu zuen, CNNak erabiliz baina moldaketa batzuk aplikatuz [132].

Kapitulu hau honela dago antolatuta: lehenik, HMMak entrenatzeko hainbat modu deskribatzen dira, alegia, zer eragin duen HMMetan datu-basearen atal desberdinak erabiltzeak, baita HMMen entrenamendu-prozesuko fase desberdinetan ere. HMM multzo bakoitzak ematen dituen PER balioak ere erakusten dira atal horretan, fonemaz fonema sailkatuta. Hurrengo atalean, azaltzen da ereduetarako parametroak nola normalizatu daitezkeen, garrantzitsua baita ereduak entrenatzeko audioaren eta azken erabiltzailearen sarrerako audioaren arteko kanal-desberdintasunarengatiko ondorioak neutralizatzea. Kanal-desberdintasunarengatiko ondorioak ikusteko, esperimentu bat aurkezten da hurrengo atalean; esperimentu horretan, mikrofono desberdinez eta distantzia desberdinetatik grabatutako audio-fitxategiak erabiltzearen ondorioak aztertzen dira. Azkenik, zenbait konklusio azaltzen dira.

5.2 HMMen entrenamendua

Ereduak egiteko hautatu den unitate fonetikoa *trifonema* ezaguna da. Trifonemak testuinguruaren araberako fonemak dira, hau da, kontuan hartzen dute aurreko eta ondorengo fonemen eragina. Esate baterako, *Aholab* hitza /*a o l a b*/[1] gisa ebakitzen da. Monofonemak erabiliz gero, lehen silabako eta azken silabako /*a*/ fonemak unitate berbera lirateke; trifonemak erabiliz gero, aldiz, desberdinak lirateke fonema bakoitzaren testuingurua desberdina da eta. Trifonemak, beraz, hiru zati ditu; hortaz, egokia dirudi hura modelatzeko hiru egoerako HMM erabiltzeak. HMMaren lehen egoerak fonemaren ezkerreko testuingurua modelatuko luke (ezkerreko fonema-trantsizioaren eskuineko erdia); bigarren egoerak fonemaren erdigunea modelatuko luke; hirugarren egoerak, berriz, eskuineko testuingurua (eskuineko fonema-trantsizioaren ezkerreko erdia). Horren azalpen grafikoa 5.1. irudian ageri da.

1 http://aholab.ehu.eus/sampa_basque.htm

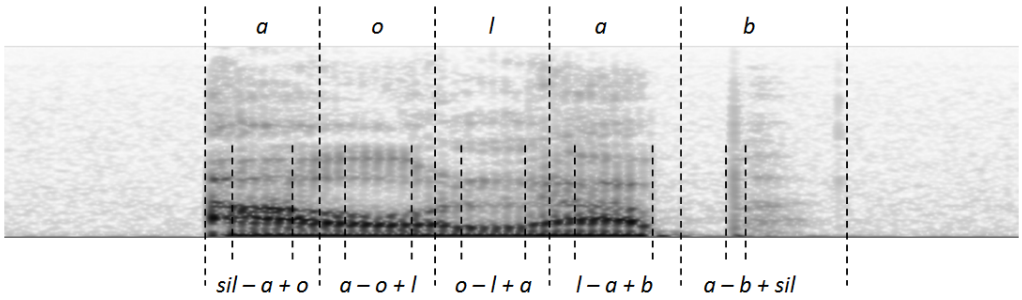


Figure 5.1: *Aholab* hitzaren errealizazio baten espektrograma eta haren zatiketa monofonematan (goian) eta trifonematan (behean).

Trifonema HMMak *HTK tool-kit* tresna [104] erabiliz sortu dira, eta, seinaleetatik parametro-bektoreak eraazteko, *AhoSR* [103] baliatu da.

HMMak entrenatzean, bi azpimultzo desberdin hartu dira kontuan (ikus datu-basearen edukia, 3.1.1. atalean):

- R azpimultzoa: Datu-basearen azpimultzo irakurria, saio bakoitzeko 25-316 elementuak dituena.
- W azpimultzoa: *Train* bloke osoa (saio bakoitzeko 1-316 elementuak).

Zatiketa hori egitearen arrazoia da ezen, arestian azaldu denez, datu-baseko azpimultzo ez irakurriak aldaera kopurua handia duela; horren ondorioz, eredu desberdinak entrenatu dira aztertzeko ea azpimultzo hori erabiltzeak merezi duen ala, bestela, zarata eransten duen.

HMMak entrenatzeko prozesuan, datu-baseko *train* bloketik bi hiztegi sortu dira, automatikoki, Aholab taldearen grafematik fonemarako transkribatzailea ($G2P$, *grapheme-to-phoneme*) erabiliz: ebakera-hiztegi kanonikoa eta ebakera alternatiboak dituen hiztegia. Hiztegi alternatibadunak aldaera fonetiko eta fonologiko dialektalak hartzen ditu kontuan, datu-basearen azpimultzo ez irakurrian agertzen diren aldaerak modelatu ahal izateko. Sarrera bakarreko hiztegiak (kanonikoak) 23 243 hitz ditu; hiztegi alternatibadunak, berriz, 95 778 ebakera desberdin ditu 23 243 lexiko-sarrera horietarako. Horrek esan nahi du *train* blokeko hitz bakoitzak, batez bestez, 4.12 ebakera desberdin dituela (lexikoiari buruzko xehetasun gehiago, 3.1.6. atalean). Horrek nahasmena eransten die eredu akustikoei, alternatibak erabiltzen baitira entrenamendu-prozesuko fase jakin batean transkripzio onena hautatu eta, hala, alternatibarik onena hautatzeko. Ebakera-alternatiba hori entrenamendu-prozesua amaitu arte erabiltzen da. Hortaz, alternatiba kopurua zenbat eta handiagoa, orduan eta handiagoa eredu akustikoetan sortutako nahasmena.

5.2.1 Lehen esperimentua

Bi HMM taldeak ebaluatzeko, ezagutze fonetikoko test bat diseinatu zen. *Test* blokeko saio bakoitzeko 3 fitxategi ebaluatu ziren (guztira, 225 fitxategi). Fitxategi horiek esaldi fonetikoki aberatsen azpimultzokoak dira, eta eskuz transkribatu ziren.

Lehen urratsa izan da R eta W azpimultzoekin sortutako HMMak ebaluatzea, ezagutze fonetikoko testak eginez. Ezagutza-probarako bilaketa-grafoa inolako murriztapenik gabe inplementatu da; hau da, fonema guztiak ekiprobableak dira, eta fonema bakoitzari beste edozeinek jarrai diezaioke. gaussiar kopuru desberdinak erabili dira, ikusteko zeinekin lortzen diren emaitzarik onenak. Bi hiztegi erabili dira: alternatibaduna eta alternatibarik gabea.

Eredu akustiko talde bakoitzarekin lortutako fonemen errore-tasak 5.2. irudian ageri dira. Errore-tasarik baxuenak 32 gaussiar erabiliz lortzen dira kasu guztietan. Gainera, hiztegi alternatibaduna erabilita ez dira emaitzak hobetzen; aitzitik, okertu egiten dira, seguruenik ebakera kopurua handia delako. 32 baino gaussiar kopuru txikiagoetarako, [R] azpimultzoaz entrenatutako ereduekin lortzen dira emaitzarik onenak; 32 baino kopuru handiagoetarako, aldiz, W azpimultzoaz entrenatutakoekin. 32 gaussiarren mugan, emaitzarik onena (% 12.74) W azpimultzoaz lortzen da hiztegi alternatibarik gabeaz; R azpimultzoaz, emaitzarik onena (% 13.35) hiztegi alternatibadunaz lortzen da.

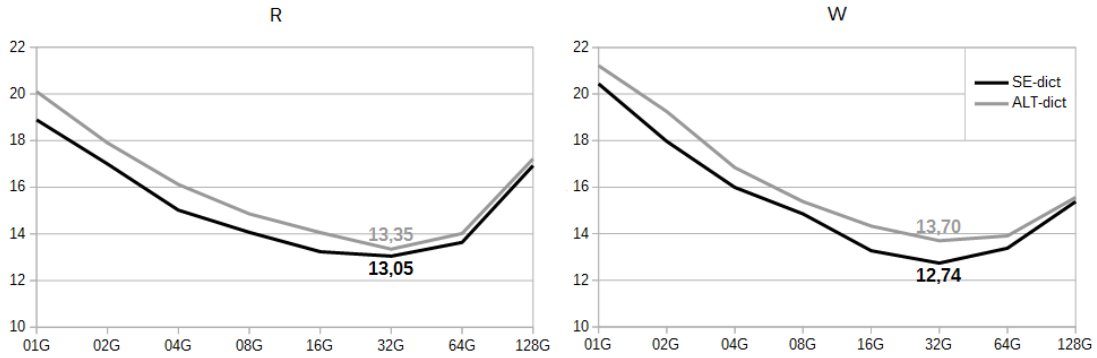


Figure 5.2: 1. esperimentua: R eta W azpimultzoez entrenatutako HMMekin ezagutza-testetan hainbat gaussiarretarako lortutako PER balioak (%), hiztegi alternatibarik gabea (SE-dict) eta hiztegi alternatibaduna (ALT-dict) erabiliz.

Test fonetikoan lortutako emaitzarik onenaren fonemaz fonemako banakapena ikusirik, zenbait konklusio atera ditzakegu. Bi metrika erabili dira fonema bakoitzaren emaitza aztertzeko: oker etiketatutako instantzien ehunekoa (E, *errors*) eta txertaketen ehunekoa (I, *insertions*), fonemaren gertaera kopuru osoarekiko (ikus 5.1. taula). Bi ratio horiek kontuan hartuta, ondorio hauetara iritsi gara:

- Bokalak, sudurkariak eta likidoak dira emaitzarik onenak dituzten fonema taldeak.
- T fonemak oso emaitza txarrak ditu bi ratioetan. Horren zergatia da hari dagokion

/T/ fonema espainierakoa delako eta ez delako oso ohikoa euskaran. Beraz, haren oso instantzia gutxi daude datu-basean —bakar batzuk hitz fonetikoki aberatsen atalean—, eta haietariko gehienak ez daude ondo ebakita.

- c ereduak ere emaitza txarrak ditu. Dagokion fonemaren soinua ez da oso ohikoa euskaran, fonema gisa —erdialdeko euskalkietan egiten da, gehienetan /t/-ren alofono gisa, oso testuinguru jakinetan—, eta, ondorioz, datu-baseko instantzia asko ez daude behar bezala ebakita. Nahasmen-matrizeak erakusten du batez ere tS ereduaz ordezkatzen dela, zeina /t/ fonemaren beste alofono bat ere baita.
- Palatalen ereduak (gj, jj eta L) nahasi egiten dira. Nahasmen-matrizeak erakusten du jj eta gj asko L gisa etiketatzen direla. Gainera, fonema beraren alofonoak dira gj eta jj, testuinguru desberdinetan ebakiak. *Grafo-kudeatzaileak* ez du kontuan hartzen bietako zein erabili behar den kasu bakoitzean.
- S ereduak E ratio eskasa du. Hari dagokion fonema ez da espainieran existitzen, eta, hortaz, datu-baseko instantzia asko ez daude ondo ebakita. Nahasmen-matrizeak erakusten du gehienetan agertzen den etiketa s dela, espainiera-hiztunek fonema hori ebakitzean egin ohi duten soinuari dagokiona.
- Fonema herskari ahoskabeen ereduek (p, t and k) E ratio onak dituzte, baina I eskasak. Horrek zergatia da ezpain-soinuetan eta hizketa ez diren soinuetan agertzen direla. Kasurik bistakoena p da.
- f ereduak oso txertaketa-ratio altua du. Gehienbat hizketa ez diren segmentu zaratatsuetan ageri dira.

Table 5.1: 1. esperimentuko ezagutza fonetikoko proban, eredu bakoitzak izandako oker etiketatutako instantzien (E) eta txertaketen (I ehunekoak.)

	a	e	i	o	u	c	p	t	k	l	r	rr	ts‘	ts	tS
E	3.9	8.7	5.6	15.2	0.9	72.3	8.5	2.7	4.2	11.9	14.4	9.7	18.4	64.1	35.1
I	13.07	16.15	16.91	13.10	10.39	51.06	58.12	26.24	23.01	18.44	13.18	35.00	41.23	14.06	16.22
	m	n	J	L	jj	gj	b	d	g	f	x	T	s‘	s	S
E	12.5	6.6	10.4	38.9	61.9	67.7	15.0	28.9	21.6	13.0	29.2	75.0	18.8	28.0	66.7
I	26.32	15.03	8.33	72.22	7.94	58.06	15.27	23.54	26.80	69.57	10.42	137.50	26.92	4.38	33.33

Arestian azaldu denez, akats askoren zergatia sistemaren hiztegiko transkripzio fonetikoen anbiguotasunak dira. Fonema batzuk modu desberdinetan ebakitzen dira; horixe da, hain zuzen, alternatibak dituen hiztegia erabiltzearen arrazoia. Hala lortutako ebakeren kopuru handiak nahasmena eragiten dio sistemari, eta emaitza fonetikoek agerian uzten dute emaitza txarragoak lortzen direla hala.

5.2.2 Bigarren esperimentua

Alternatiba desberdinen artetik transkripzio zuzena hartzeko konponbide bat da transkripzio kopuru txiki bat eskuz zuzentzea. Horrek aukera ematen du HMMak hasierako fasean behar bezala lerrokatzeko eta, hala, hurrengo faseetan erroreak gutxiagotzeko. Hasieran 12 eta gero 25 saio zuzendu ziren eskuz (saio osoak). Zuzentzean, transkripzio fonetikoak nahiz gertaera akustikoen etiketak hartu ziren kontuan. Zuzendutako saioak datu-baseko lehen 12 saioak (hemendik aurrera: *M12*) eta lehen 25 saioak (hemendik aurrera: *M25*) izan ziren.

Horiek horrela, bigarren proba bat egin zen, ikusteko zein diren *M12* eta *M25* erabiltzearen ondorioak. Bigarren proba honetan, HMM berriak garatu dira R eta W azpimultzoekin batera eskuz zuzendutako saioak erabiliz, eta ezagutza-proba berri bat gauzatu da. 5.3. irudian, *R+M12* testean erabili diren azpimultzo desberdinak ageri dira, adibide gisa. Kontuan izan saio bakoitzari hizlari bat dagokiola.

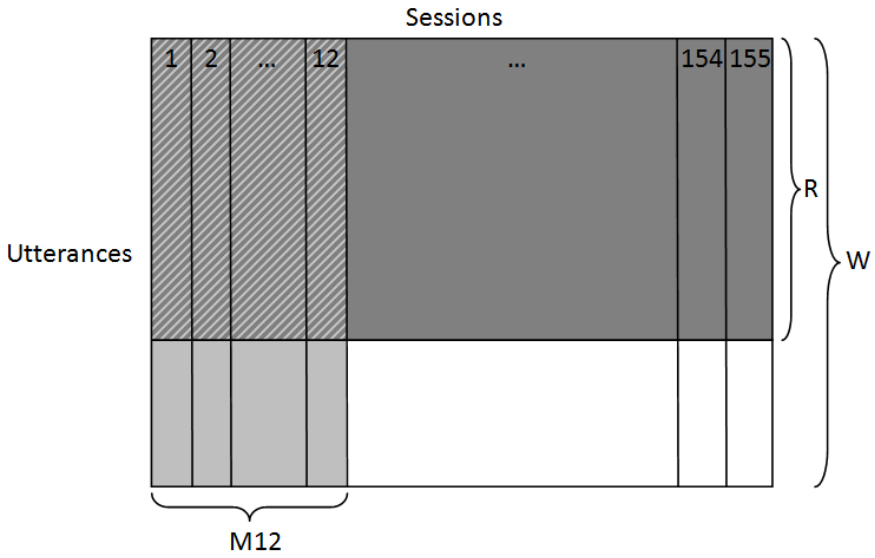


Figure 5.3: *Basque Speecon-like* datu-baseko *train* blokeko azpimultzoen adibidea, *R+M12* erabiliz HMMak entrenatzeko.

Testatutako fitxategiak aurreko esperimentuan erabilitako fitxategi berak dira. 5.4. irudian, gaussiar kopuru desberdinetarako lortutako PER balioak ageri dira, 5.2. irudiko datuekin batera (beltzez), emaitzen ikuspegi hobea izateko.

Emaitzek azaltzen dute errore-tasak, espero bezala, baxuagoak direla eskuz zuzendutako saio gehiago kontsideratu ahala. Test honetan, ia kasu guztietan, emaitzarik onenak R azpimultzorako (eta 32 gaussiarretarako) lortzen dira. Hala ere, esperimentu honen helburua zen HMMen sorreran lerrokatze hobea lortzea, eta emaitzarik onenak, oraindik ere, hiztegi alternatibarik gabeaz lortzen dira. Dena dela, hiztegi alternatibarik gabearen probako emaitzarik onena (% 12.34) eta hiztegi alternatibadunaren probako

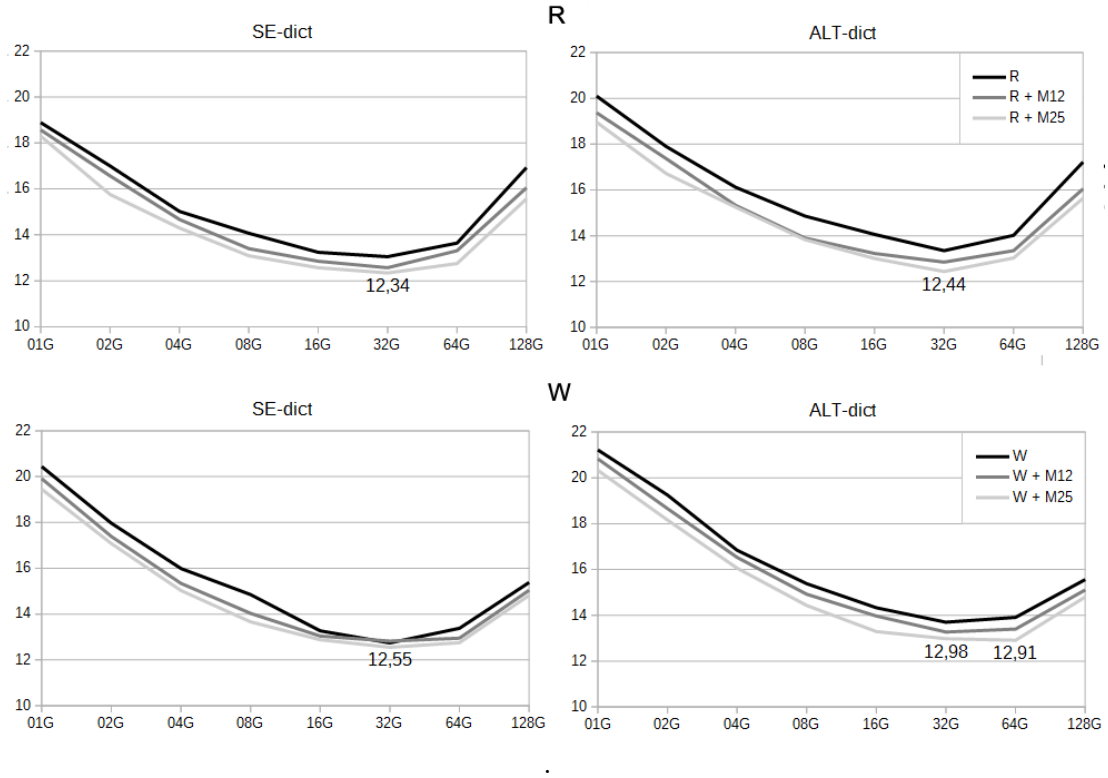


Figure 5.4: 2. esperimentua: *R* (goian) eta *W* (behean) azpimultzoez entrenatutako HMMekin ezagutza-testetan hainbat gaussiarretarako lortutako PER balioak (%), eskuz zuzendutako saiorik gabe eta saioekin (—, *M12* eta *M25*), hiztegi alternatibarik gabea (SE-dict) eta hiztegi alternatibaduna (ALT-dict) erabiliz

emaitzarik onena (% 12.44) hurbilago daude orain. Dirudienez, eskuz zuzendutako saioak erabiltzeak eragin handiagoa du hiztegi alternatibadunarekin egindako esperimentuetan; orobat, eragin handiagoa du *R* azpimultzoarekin entrenatutako HMMetan.

5.2.3 Hirugarren esperimentua

Kontuan izanda helburu nagusia dela HMMak sortzean ebakera alternatiboen lerrokatze egokia lortzea, hirugarren proba bat egitea pentsatu da. Hemen, HMMak sortzeko hainbat modu kontsideratu dira. HMMak entrenatzeko prozesuak zenbait fase ditu, monofonemetatik hasi eta trifonemak sortzeraino. Eskuarki, hasierako faseetan, hiztegi alternatibarik gabea erabiltzen da, eta, gero, erdiko fase batean, lerrokatze behartuzko ezagutze-prozesu bat gauzatzen da hiztegi alternatibaduna erabiliz, hitz bakoitzaren transkripzio fonetikorik onena hautatzeko. Fase horretatik aurrera, lerrokatze-prozesutik ateratako transkripzioak erabiltzen dira, eta, hala, bermatzen da hitz bakoitzaren transkripzio alternatiborik onena erabiltzen dela. Hasierako HMM horiek nola estimatzen diren, beraz, erabakigarria da transkripzio-lerrokatze onak lortzeko; hasierako

transkripzioak zenbat eta zehatzagoak, orduan eta hobeak azken HMMak.

Aurrekoa kontuan hartuta, hirugarren testa honetan datza: entrenamendu-prozesuko lehen faseetarako eta hurrengoetarako, azpimultzo desberdinak baliatu dira (haietan guztietan hiztegi alternatibarik gabea eta hiztegi alternatibaduna erabiliz). 5.5. irudian, *M25*ren emaitzak ageri da (*M12*renak baztertu egin dira). Aipatzekoa da erabilitako nomenklatura: "*M25* — *R+M25*"ek esan nahi du entrenamendu-prozesuko lehen faseetan lehen 25 saioak soilik erabili direla (*M25* azpimultzokoak), eta, gero, transkripzio-lerrokatzearen ondoren, *R* eta *M25* azpimultzoen bildura.

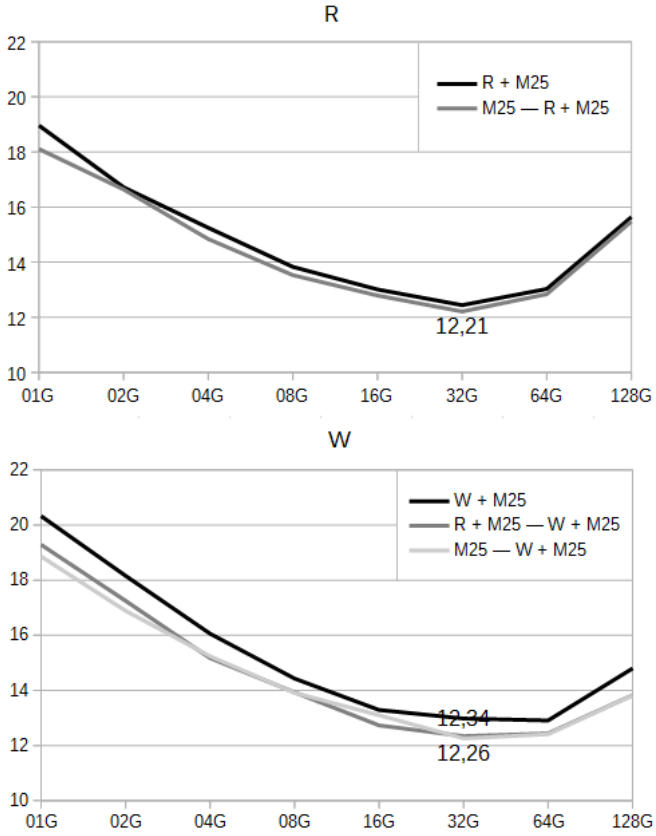


Figure 5.5: 3. esperimentua: Faseka entrenatutako HMMekin —faseka ez direnen aldean (beltzez)— ezagutza-testetan hainbat gaussiarretarako lortutako PER balioak (%), hiztegi alternatibaduna erabiliz (*M25*erako).

Fasekako entrenamendu-teknika horrekin, orain arteko emaitzarik onenak lortu dira: % 12.21 eta % 12.26, hurrenez hurren *R* ("*M25* — *R+M25*") azpimultzoaren eta *W* ("*M25* — *W+M25*") azpimultzoaren esperimentuei dagozkienak. Gainera, % 12.34ko emaitza (2. esperimentuko emaitzarik onenaren berdina) lortu da "*R+M25* — *W+M25*" moduan ere. Nahiz eta 2. testeko emaitzarik onenaren (hiztegi alternatibarik gabea erabiliz) eta

3. test honen arteko aldeak oso txikiak izan, bistan da azkenik lerrokatze hobea lortu dela hiztegi alternatibaduna erabiliz eta hasieran HMMak oso transkripzio-errore gutxi dituen azpimultzo batez eta, gero, azpimultzo handiago batez entrenatuz.

Hirugarren esperimentu honetan lortutako emaitzak hobeto ulertzeko, bi grafiko argigarri daude 5.6. irudian. Hiztegi alternatibaduna erabiliz hiztegi alternatibarik gabearekiko lortzen den aldea (balio absolutua) ageri da bi grafikoetan, eskuz zuzendutako transkripzio kopuru desberdinetarako (*0*, *M12* edo *M25*). Ezkerreko grafikoan, *R* azpimultzoa erabiliz lortutako emaitzak ageri dira; eskuinekoan, berriz, *W* azpimultzoa erabiliz lortutakoak. Grafiko bakoitzean bi grafiko daude: ezkerrekoa, faseka entrenatu ez diren HMMen emaitzei dagpkie; eskuinekoa, berriz, faseka entrenatutako HMMen emaitzei.

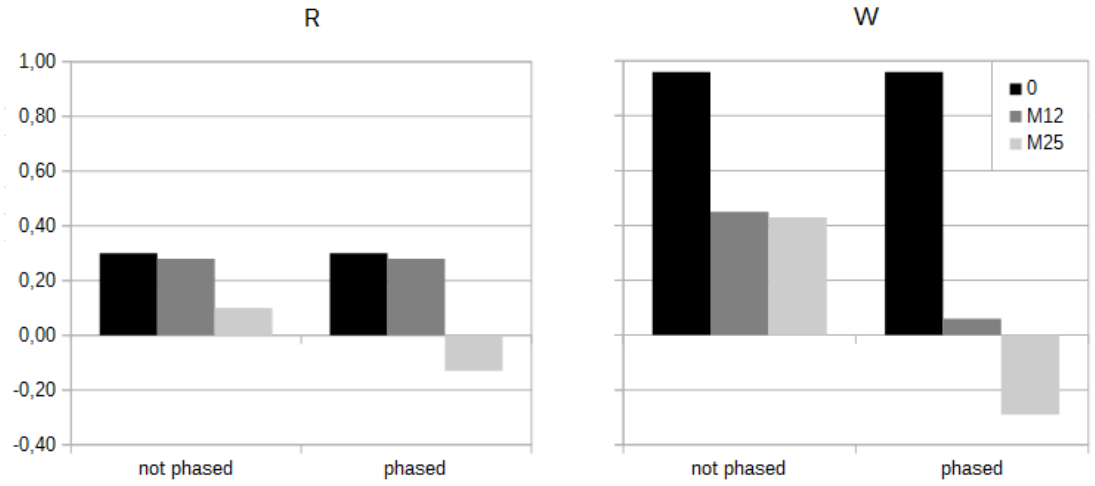


Figure 5.6: Hiztegi alternatibaduna eta faseka eta faserik gabe entrenatutako HMMak (32 gaussiarrekoak) erabiliz lortutako PER diferentzia absolutuak hiztegi alternatibarik gabeaz egindako emaitzekiko, eskuz zuzendutako hainbat transkripzio kopurutarako (ezkerrean, *R* azpimultzoaren emaitzak; eskuinekoan, *W* azpimultzoarenak).

Diagramek erakusten dute ezen, bai *R* azpimultzorako ("*M25* — *R+M25*"), bai *W* azpimultzorako ("*M25* — *W+M25*") faserik gabe entrenatutako ereduekin alternatibak erabiliz ez direla emaitza hobeak lortzen hiztegi alternatibarik gabea erabiliz lortutakoen aldean (nahiz eta, espero bezala, eskuz zuzendutako transkripzio kopurua zenbat eta handiagoa, orduan eta hobeak emaitzak). Hala ere, faseka entrenatutako ereduekin alternatibak erabiliz, emaitza hobeak lortzen dira; *M25*erako, hobekuntza ere lortzen da. Aipatu beharra dago eskuz zuzendutako transkripziorik gabe lortutako emaitzak (barra beltzak diagrametan) berdinak direla fasekako eta faserik gabeko modalitateetan, azpimultzo bera erabiltzen baita entrenamendu-fase bakoitzean.

Faseka entrenatutako ereduak erabiltzeak fonema-mailako emaitzetan duen eragina ikusteko, fonemaz fonemako banakapena ere sortu da emaitzarik onenak izan dituzten HMMentzat ("*M25 — R+M25*", hiztegi alternatibaduna eta 32 gaussiar). Emaitzak 5.2. taulan ikus daitezke.

Table 5.2: Oker etiketatutako instantzien (E) eta txertaketen (I) ehunekoak fonemaka, 3. esperimentuan emaitzarik onena izan duen ezagutze fonetikoko testean.

	a	e	i	o	u	c	p	t	k	l	r	rr	ts‘	ts	tS
E	4.1	7.5	6.5	5.8	11,3	61.7	11.1	3.6	4.1	13.7	17.8	7.5	16.1	57.6	35.1
I	10.57	13.96	14.03	10.98	8.75	10.64	41.03	15.74	18.48	13.71	6.02	22.96	28.57	15.15	18.92
	m	n	J	L	jj	gj	b	d	g	f	x	T	s‘	s	S
E	15.0	6.8	10.4	55.6	54.8	33.5	16.5	17.6	20.6	4.3	6.2	62.5	21.2	18.9	58.3
I	20.26	10.41	29.17	55.56	130.65	6.45	11.85	16.84	26.60	47.83	11.46	137.50	15.87	4.84	50.00

Fonema-mailako emaitzak hobeto interpretatzeko, konparazio-irudi bat sortu da: 5.7. irudia. Irudian, 5.1. taulako emaitzak (barra grisak) eta 5.2. taulakoak (barra beltzak) ageri dira, ereduz eredu: goiko grafikoan E ratioa ageri da, eta behekoan, I ratioa. Oro har, E ratio txikiagoak lortzen dira 3. testeko HMMez, nahiz eta zenbait kasutan (u eta L ereduak) E nabarmen igotzen den. Iari dagokionez, oro har emaitza hobeak lortzen dira, jj ereduaz izan ezik, nabarmen igo baita haren txertaketa-ratioa.

Lerrokatze desegoki baten ondorioa da jj ereduaren arazoa. Euskarako zenbait dialektotan, /i/z amaitutako izen eta adjektiboei soinu gehigarri bat eransten zaie -a artikulua eranstean (bien artean jartzen da soinu gehigarria). Soinu desberdinak txertatzen dira euskalkiaren arabera (/jj/ erdialdeko euskalkiei dagokie), eta aldaera guztiak daude hiztegi alternatibadunean. Hala ere, eskuz zuzendutako azpimultzoetan, halabeharrez, ez da behin ere ageri, eta deskodetzaileak ez du datu nahikorik hitzak aldaeradunak ala aldaerarik gabeak diren erabakitzeko. Hortaz, lerrokatze-prozesuan okerreko fonema asko txertatu dira. Eskuz zuzendutako datu dialektal kopuru txiki bat nahikoa izango da arazoa konpontzeko. Dena dela, etorkizunean garatzeko utzi da lan hori, trifonema jakin bati besterik ez baitio eragiten horrek.

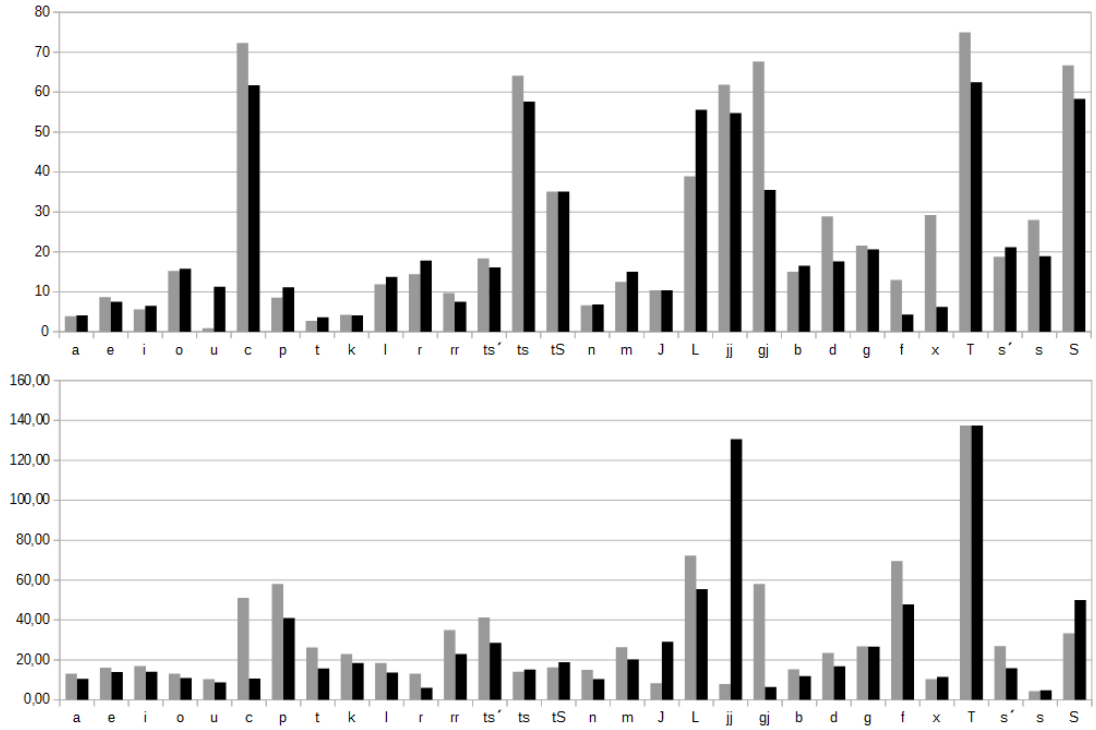


Figure 5.7: 1. esperimentuan (zutabe grisak) eta 3.ean (zutabe beltzak) emaitzarik onenak izandako HMMen ezagutza fonetikoen emaitzen konparaziozko irudia. Goian, oker etiketatutako instantziak (E); behean, txertaketak (I).

5.3 Parametro-normalizazioa: CMVN

Parametro-normalizazioa ezinbestekoa da eredu akustiko sendoak sortzeko eta mikrofono desberdinekin jasotako audio-seinaleekin aritzeko. Lehendabiziko esperimentuak, 6. kapituluan azaldu bezala, PC berean egin dira ikasle guztiek entzungailu mikrofonodun berbera erabilita (parametro-normalizaziorik gabe); baina sistemaren web inplementazioak nolabaiteko normalizazioa behar du, ikasle bakoitzak bere ekipamendua erabiliko baitu.

Sarrerako seinaleen arteko desberdintasunak (kanalak, hondoko zarata eta abar) konpentsatzeko, metodorik ohikoena da erauzitako parametroei Batezbesteko eta Bariantza Normalizazio Cepstrala (CMVN, *Cepstral Mean and Variance Normalisation*) aplikatzea. Tesi honetan ere metodo hori hautatu da, eta, beraz, eredu akustiko berriak entrenatu eta testatu dira MFCC normalizatuak erabiliz. CMVNren oinarriari buruzko xehetasun gehiago izateko, jo 9. kapitulura.

[133]n azaltzen denez, MFCCen batezbesteko-bektoreak uneko mikrofonoaren eta gelaren akustikaren ezaugarri espektralak hartzen ditu kontuan. Limitean, grabazio bakoitzean $N \longrightarrow \infty$ joanez gero, espero izatekoa litzateke grabazio-ingurune bereko parametro akustikoen batezbestekoak berdinak izatea. *Basque Speecon-like* datu-baseko

hizlari bakoitza kondizio akustiko berean grabatu denez, datu-baseko audio-seinaleei dagokien MFCCak hizlarika normalizatu dira. Batezbesteko- eta bariantza-bektoreak kalkulatu dira hizlari bakoitzaren audio guztiak erabiliz (316 fitxategi). Ondoren, normalizatutako MFCC fitxategiekin, HMM berriak entrenatu dira.

5.3.1 Esperimentuak

Teknika honen eragina ikusteko, aurreko atalean (5.2. atala) azaldutako hiru esperimentuak berriro egin dira, HMM normalizatuak erabiliz. Espero da normalizazioaren ondorioz emaitza onak lortzea, testatutako fitxategiek desberdintasunak dituztenean, HMMak garatzeko erabilitako fitxategiekin alderatuta, audioa jasotzeko ekipamenduari, inguruneari edo hondo-zaratari dagokionez. Testatutako fitxategiak entrenamenduko datuen datu-base berekoak direnez, testuinguru honek ez dirudi onena denik normalizazioaren emaitzak behar bezala ikusteko. Hala ere, HMM normalizatuen jokaera ulertzeko lagungarria izan daiteke. Hiru test berrietan lortutako emaitzak irudi hauetan bildu ditugu: 5.8. irudian (1. esperimentua), 5.9. irudian (2. esperimentua) eta 5.10. irudian (3. esperimentua).

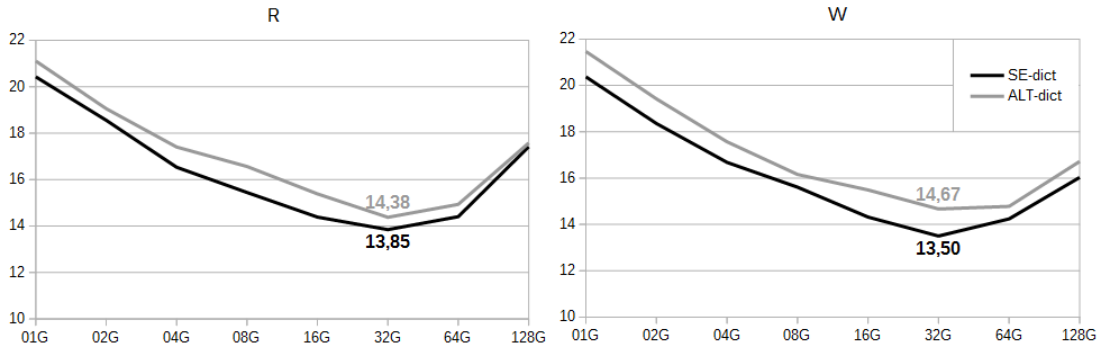


Figure 5.8: 1. esperimentua CMVN aplikatuz berregina.

Parametroak normalizatu gabe lortutako emaitzak baino pittin bat okerragoak dira parametroak normalizatuta lortutako emaitzak. Denetan emaitzarik onena, % 12.45ekoa, *R+M25* azpimultzoaz lortzen da, hiztegi alternatibarik gabeaz. Gainera, ematen du mugitu egin direla gaussiar kopuruetan zeharreko kurbak; izan ere, esperimentu gehiagok dute orain minimoa 64 gaussiarretan, batez ere *W* azpimultzokoek.

Gainera, *M12* eta *M25* azpimultzoak soilik erabiliz ere egin ditugu esperimentuak. Orain arte ez ditugu kontuan hartu, zeren, guztiz etiketatuta eta zuzen badaude ere, ez dute HMMak modu egokian entrenatzeko nahikoa daturik. Horrenbestez, esperimentu fonetiko berak egin dira *M12* eta *M25*rako, CMVNrik gabe eta CMVNarekin, eta, orain arte gertatutakoaren kontrara, emaitzak hobeak dira CMVNarekin. 5.11. irudian, grafikoki ageri dira bi kasu horietarako lortutako emaitzak, *R*, *R+M12* eta *R+M25* azpimultzoekin batera, azpimultzo desberdinak erabiliz entrenatutako HMMen jokaera desberdinak ikustearren.

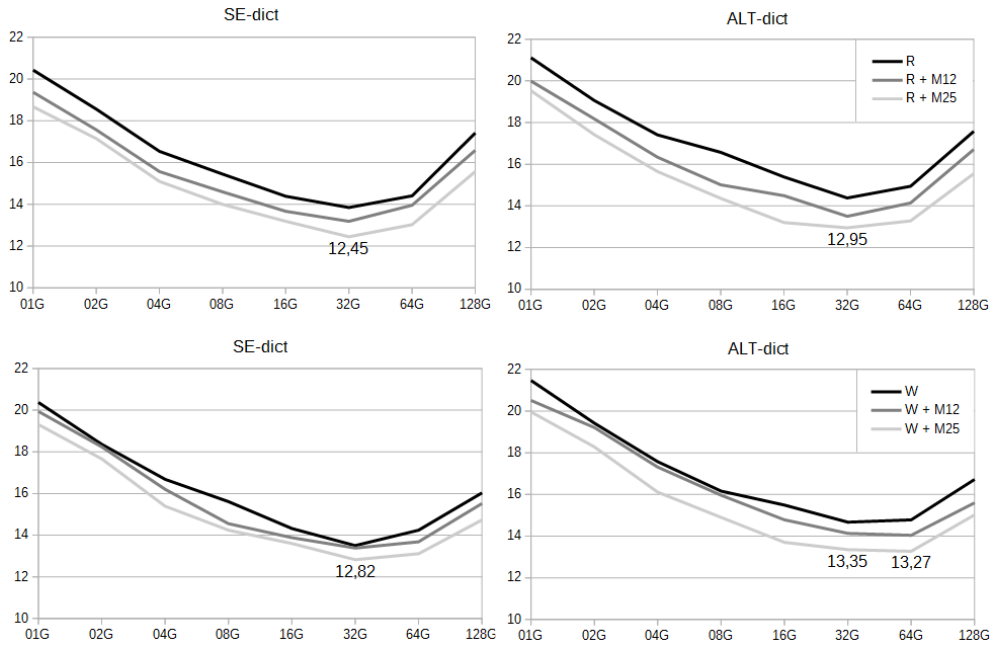


Figure 5.9: 2. esperimentua CMVN aplikatuz berregina.

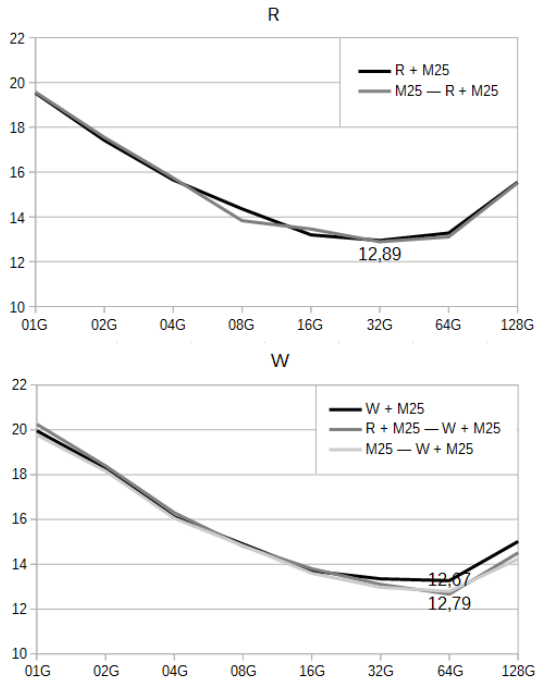


Figure 5.10: 3. esperimentua CMVN aplikatuz berregina.

R azpimultzoarentzat antzeko emaitzak eta *R+M12* eta *R+M25* azpimultzoetarako emaitza okerragoak lortzen badira ere, % 1eko hobekuntza absolutua lortu da *M12* eta *M25* azpimultzoetarako. Horrek esan nahi du parametroen normalizazioa onuragarria izan daitekeela transkripzioetako errore kopurua txikia denean.

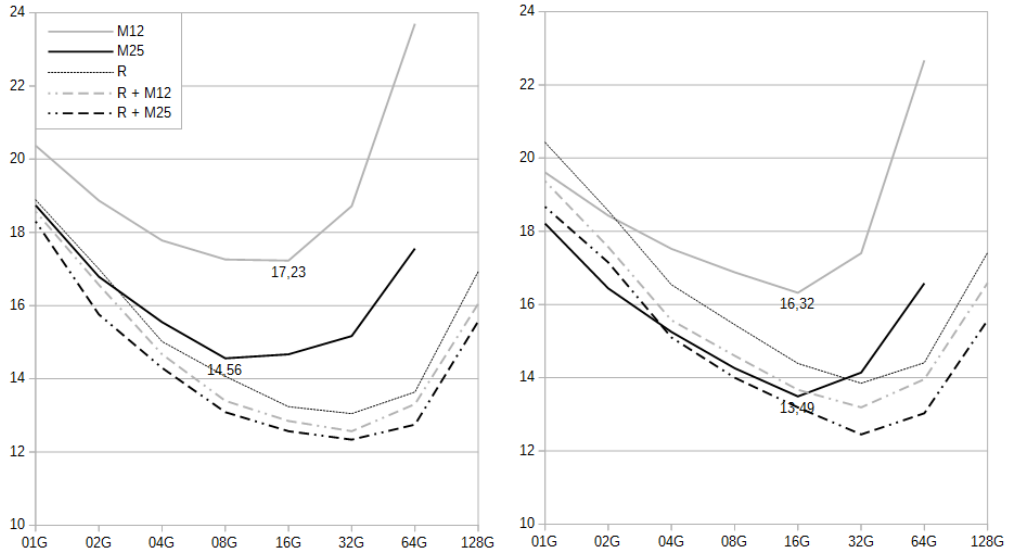


Figure 5.11: *M12* eta *M25* azpimultzoak erabiliz entrenatutako HMMen fonemen errore-tasa (%), beste azpimultzo batzuekin batera CMVNrik gabe (ezkerrean) eta CMVNarekin (eskuinean).

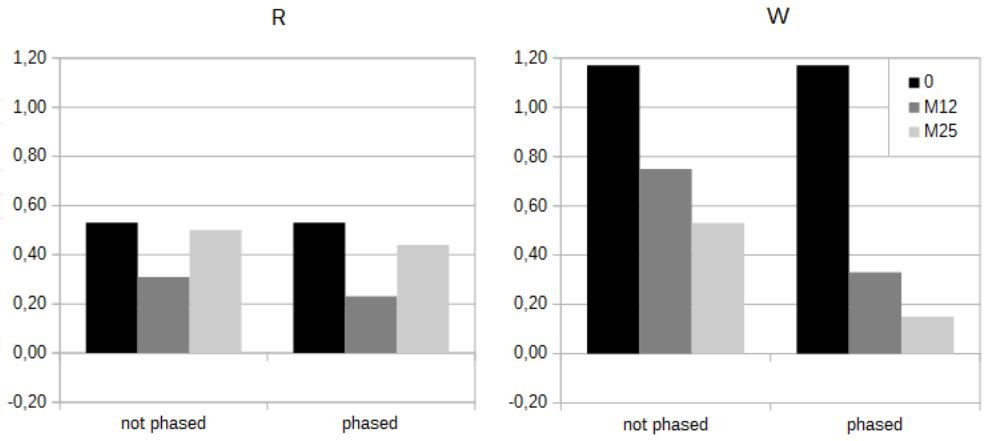


Figure 5.12: Hiztegi alternatibaduna eta faseka eta faserik gabe entrenatutako HMMak (32 gaussiar) erabiliz egindako CMVN esperimentuen PER diferentzia absolutuak, hiztegi alternatibarik gabearekiko, eskuz zuzendutako transkripzio kopuru desberdinetarako (ezkerrean: *R* azpimultzoa erabiliz lortutako emaitzak; eskuinean: *W* erabiliz).

CMVNaren emaitzetan, hiztegi alternatibaduna erabiliz eta hiztegi alternatibarik gabea erabiliz lortzen diren emaitzak alderatuta, 5.6. irudiaren antzeko irudi bikoitza sortu da (ikus 5.12. irudia). 5.6. irudian ez bezala, hiztegi alternatibaduna erabiliz ez dira lortzen hiztegi alternatibarik gabea erabiliz lortutako emaitzak baino hobeak, nahiz eta emaitzak hobeak izan eskuz zuzendutako transkripzio gehiago erabili ahala. Nabarmentzekoa da ezen errore-igoera, R azpimultzorako, handiagoa dela *M25*erako, *M12*rako baino. Hala eta guztiz ere, emaitzak hobeak dira *M25*erako.

5.4 Kanal-desberdintasuna testatuz

CMVNaren ondorioak ez dira oso nabariak, testatutako seinaleen grabazio-baldintza akustikoak eta HMMak entrenatzeko erabilitako seinaleenak antzekoak baldin badira. Gure tresna zerbitzari batean kokaturik dagoenez eta erabiltzaile bakoitzak mikrofono desberdin bat, ingurune desberdin batean, erabiliko duenez, kanal-desberdintasunak egongo dira entrenamenduko datuen eta deskodetzaileak jasoko dituen seinaleen artean. Suposatzen da balizko egoera horretan nabariak izango direla normalizazio cepstralaren abantailak.

Kontu hori aztertzeko, beste test bat egin da: aurreko atalean deskribatu bezala entrenatutako HMMak (bai normalizatu gabeak, bai normalizatuak) erabili dira, grabazio-kanal desberdina duten audio-seinaleak testatzeko. Xede horretarako, *Basque Speecon-like* datu-baseko *mahai gaineko* mikrofonoaren bidez (1 m-ko distantziara) grabatutako blokeko fitxategiak erabili dira. Test-fitxategi berri horiek eta lehenago erabilitako fitxategiek (225 fitxategi, hizlariko 3 fitxategi —ikus 5.2.1. atala—) eduki bera dute; desberdintasun bakarra da mahai gaineko mikrofono batez grabatuak direla, hurbileko entzungailu batez grabatuak izan beharrean.

5.3. taulan, test horietan lortutako emaitza batzuk ageri dira (PER (%)). Nahiz eta aurreko testetan modu desberdinetan entrenatutako HMM guztiak erabili diren, taulan erakusteko daturik adierazgarrienak hautatu dira: 32 gaussiarrekin eta eskuz

Table 5.3: *Basque Speecon-like* datu-baseko *mahai gaineko* azpicorpusaz egindako ezagutza fonetikoko testen Fonema Erroreen Tasak (%), *hurbileko* azpicorpusaz entrenatutako HMMak erabiliz, CMVNaz eta CMVNrik gabe (32 gaussiar).

			CMVNrik gabe	CMVNarekin
SE-dict	*R*	*M25*	26.56	21.87
	W	*M25*	**25.92**	**21.83**
ALT-dict	*R*	*M25*	27.21	22.41
	W	*M25*	26.37	22.55
	M25 — *R+M25*		27.32	22.15
	M25 — *W+M25*		26.93	22.09
	R+M25 — *W+M25*		27.18	21.85

zuzendutako 25 saio erabiliz (*M25*) entrenatutako HMMekin egindako testa. Fitxategiak CMVNaz eta CMVNrik gabe testatu dira.

Emaitzek kanal-desberdintasunaren eragina agerian jartzen dute. Normalizaziorik gabe, lortzen den PER onena % 25.92 da, *mahai gaineko* azpicorpusa testatuz (kanal desberdinak); hala ere, % 12.21 lortzen da *hurbileko* fitxategiak testatuz (kanal berdinak). Beraz, kanal desberdin batekin testatuz lortzen diren fonema-erroreen tasak kanal berdinarekin lortzen dena halako bi baino pittin bat handiagoak dira. CMVNarekin, lortzen den PER onena % 21.83 da *mahai gaineko* fitxategiak testatuz, eta % 12.45 *hurbileko* fitxategiak testatuz. Kasu horretan, fonema-erroreen tasa, kanal desberdinen baldintzapean, ez da bikoitzera iristen; nahiko altua da, halere. Dena dela, % 17ko hobekuntza orokorra lortu da CMVNarekin, kanal desberdinen baldintzapean egindako test guztiak kontuan hartuta; kanal berdineko seinaleak testatzean, berriz, ia ez da alderik sumatzen.

5.5 Konklusioak

Atal honetan, hainbat modutan sortutako HMMak entrenatu eta testatu ditugu. Ikusi dugu ebakera-hiztegi alternatibaduna erabiltzeak okerreko eragina duela, zeren ebakera alternatiboen kopuru handia hartu behar baita kontuan. Hori dela eta, fasekako HMM entrenamenduari ekin zaio. Helburua da ebakera-lerrokatze ona lortzea hiztegi alternatibaduna erabiliz, eta, horretarako, eskuz zuzendu da datu-basearen zati txiki bat, lerrokatzea egokia izateko gida gisa balioko duelakoan. Zuzenketa horiek, fasekako entrenamenduarekin batera, emaitzarik onena lortu dute: % 12.21eko PER ("*M25 — R+M25*" prozesua, hiztegi alternatibaduna erabiliz eta 32 gaussiar). Emaitza onak lortu badira ere, hobekuntza handiagoa espero zen hasiera batean. Esperotako hobekuntza lortu ez izanaren arrazoietako bat jj fonemaren txertaketa-tasa handia da; izan ere, ez da oso fonema ugaria euskaraz, eta ez dago behar bezala entrenatuta, fonema horren agerpenik ez baita ageri eskuz zuzendutako azpimultzoetan.

Parametro normalizatuak erabiliz entrenatutako HMMak testatzean, ia emaitza berdinak lortu dira (pittin bat okerragoak), seguruenik normalizazioaren ondorioak ezin direlako ikusi ekipo berarekin eta, gehienetan, ingurune berean grabatutako fitxategiak testatzean. Hala ere, CMVN erabiltzearen onurak bistakoak dira eskuz zuzendutako datuak bakarrik erabiliz egindako ezagutza fonetikoko testean.

Gainera, test guztiak errepikatu egin dira, ikusteko nolakoa den HMMen jokaera ingurune desberdinetan eta mikrofono desberdinak erabiliz grabatutako audio-fitxategiak prozesatzen direnean. Horretarako, 1 m-ko distantzia batera ezarritako mikrofono baten bidez grabatutako fitxategiak testatu dira. CMVNrik gabe, lortzen den errore-tasaren igoera % 100 baino handiagoa da. CMVNarekin, errore-tasaren igoera txikiagoa lortzen da, % 70 ingurukoa. Hala ere, CMVNak eragin handia du emaitzetan; ez espero bezain handia, halere.

Emaitzak aztertuta, bi HMM multzo hautatu dira azken sistemaren diseinurako:

- **CMVNrik gabe**: "*M25 — R+M25*" prozesuari jarraiki sortutako HMMak

kontsideratu dira, hiztegi alternatibadunaz eta 32 gaussiarrez.

- **CMVNarekin**: "*R+M25*" prozesua jarraiki sortutako HMMak, hiztegi alternatibarik gabeaz eta 32 gaussiarrez.

Aipagarriak dira, halaber, kapitulu honetan lortu diren PERak, nabarmen txikiagoak baitira gaur egun TIMIT datu-baserako lortzen dena baino (% 16.5, kapitulu honen sarreran azaldu bezala), puntako teknikak erabili gabe ere. Logikoki, bi datu-baseen arteko desberdintasunek eragin handia dute emaitzetan.

CHAPTER 6

Hastapenetako esperimentuak eta hasierako sistema

6.1 Sarrera

Lan honen hasieran, ez genuen kontsideratu ere egin sistema Interneteko zerbitzari batean inplementatzeko aukera. Audioa jasotzeko tresnak ez zeuden nabigatzaileetan inplementaturik, eta bateragarritasun-arazo handiak ematen zituzten *Flash*, *Java applet* edo antzeko plugin edo kanpo-aplikazioek. Horren ondorioz, lokalean exekutatzeko diseinatu zen, hasiera batean, *AhoSR*.

AhoSR-ren lehendabiziko prototipoa 2009an zegoen erabilgarri, baina 2011n ekin genien hizkuntzak ikasteko lehen egiaztapen-esperimentuei. Urte hartan, gure *AhoSR* sistemak ez zeukan, ez ahots-aktibitatea detektatzeko sistemarik (VAD, *Voice Activity Detection*), ez batezbesteko- eta bariantza-normalizazio cepstralik (CMVN, *Cepstral Mean and Variance Normalisation*) (ezaugarri horiek tesi honen testuinguruan garatu dira). *Basque Speecon-like* datu-baseak, HMMak entrenatzeko erabili zenak, ez zuen transkripzio fonetikorik berez, eta lan luzea aurreikusten zen arazo horri aurre egiteko.

Testuinguru horretan, ASR-rako erabiltzen ari ginen ereduak datu-basearen R azpimultzoa (ikus 5.2. atala) erabiliz sortutako HMMak ziren, alegia, bat-bateko hizketa duten fitxategiak alde batera utzita. Datu asko baztertu baziren ere, lehendabiziko esperimentuak diseinatu ziren R azpimultzoaz, aldaera dialektalak kudeatzeko modu bat aurkitu bitartean.

Lan guzti honen hasiera-puntu gisa, saiatu ginen ikusten ea ASR-rako sortutako HMMak erabilgarriak ziren fonemak bereizteko konfiantza-puntuazioak entrenatzeko (adibidez, GOPak). Lehendabiziko esperimentuak laborategian egin ziren hasieran, baina emaitzak mundu errealean lortzeko beharra ere ikusi zen. Hala, *AhoSR_L2* izeneko softwarea garatu genuen C++ lengoaian, lokalean exekutatzen zena.

Kapitulu honetan, sarrerako fonema bat zuzen ala oker ebakita dagoen ala ez balioesteko hautatu genuen metodoa aurkezten da, GOP puntuazioetan oinarritua. Berez, *Basque Speecon-like* datu-basea ez da OBEL datu-base bat, eta, hortaz, ez du grabaziorik bigarren hizkuntzen hiztunek oker ebakitako fonemak dituenik. Hala, ebakera okerreko fonemen GOP banaketa lortzeko metodo baten beharra sentitu genuen. Metodo

hori hurrengo atalean azaldurik dago (6.2. atala). 6.3. atalean, ebakera ebaluatzeko atazaren lehendabiziko ebaluazioa deskribatzen da, laborategiko baldintzetan egina. Ondoren, *AhoSR_L2* softwarearen deskripzioa ageri da 6.4. atalean. Hizkuntza-ariketen diseinua eta ebaluazioa aurkezten da gero 6.5. atalean, eta, azkenik, zenbat konklusio azaltzen dira.

6.2 Fonema-puntuazioa: GOPak eta erabaki-atalaseak

6.2.1 Oinarrizko GOP algoritmoa

q_i fonemaren GOP puntuazioa lortzeko, bere $p(q_i|O_i)$ ondorengo probabilitatearen logaritmoa kalkulatzen da, iraupenaz normalizatuta, O_i segmentu akustikoan zehar (i: hitzaren barneko fonema-indizea), (6.1) ekuazioan zehazten denez [134].

$$GOP\left(q_i\right) = \frac{1}{T_i} logp\left(q_i|O_i\right) = \frac{1}{T_i} log\left[\frac{p\left(O_i|q_i\right)p\left(q_i\right)}{\sum_{j=1}^{N} p\left(O_i|q_j\right)p\left(q_j\right)}\right] \tag{6.1}$$

non T_i segmentuak dirauen bilbe kopurua den eta N fonema kopuru osoa den. Jotzen badugu fonema guztiak ekiprobableak direla ($p(q_i) = p(q_j)$) eta izendatzaileko batura bere maximoaz ordezkatu daitekeela, GOP neurria honela geratzen da:

$$GOP\left(q_i\right) \approx \frac{1}{T_i} log\left[\frac{p\left(O_i|q_i\right)}{p\left(O_i|q_{j_{max}}\right)}\right] \tag{6.2}$$

non j_{max} segmentu horretarako egiantzik altuena ematen duen fonema-ereduaren indizea den. Segmentu akustikoaren mugak eta hari dagozkion egiantzak Viterbi lerrokatzetik erauzten dira. Lehendabizi, (6.2) ekuazioko zenbakitzailea kalkulatzen da lerrokatze behartu moduan, ezaguna izanik fonema-ereduen sekuentzia. Bigarrenik, izendatzailea kalkulatzen da murriztapenik gabeko fonema-begizta bat baliatuz. Hala, izendatzaileko puntuazioa kalkulatzen da, bilbeko log-egiantzak batuz, $O_i$ren segmentuan zehar. Praktikan, horrek esan nahi du maiz murriztapenik gabeko fonema sekuentziako fonema batek baino gehiago parte hartzen duela $p(O_i|q_{j_{max}})$ kalkulatzean.

6.2.2 Atalaseak ezarriz

Teorian, erabakitzeko ea fonema bat zuzen ala oker ebakita dagoen, GOP puntuazioak erabiliz, bi GOP banaketa behar dira: batetik, zuzen ebakitako instantzien GOP banaketa, edozein ASR datu-basetatik lor daitekeena (gure kasuan, *Basque Speecon-like* datu-basea), GOPak lerrokatze behartu moduan kalkulatuz. Bestetik, oker ebakitako instantzien GOP banaketa. Azken datu horiek lortzea zailagoa da.

Literaturan, hainbat modu deskribatu izan dira, oker ebakitako fonemen GOP atalaseak kalkulatzeko:

1. **Kalkulu enpirikoa**: entrenamendu-datuetan dauden q_i fonemen GOP guztien μ_i batezbestekoen eta σ_i bariantzen arabera jar daiteke q_i fonema baten atalasea, (6.3) ekuazioan ageri den moduan:

$$T_{q_i} = \mu_i + \alpha\sigma_i + \beta \tag{6.3}$$

   non α eta β enpirikoki zehazten diren eskalatze-konstanteak diren. [46]n, balio hauek ematen zaizkie: $0.8 < \alpha < 1.3$ eta $-1.0 < \beta < 2.0$; horrekin, atalase orokorren antzeko eskalan ateratzen dira emaitzak, baina fonema bakoitzera egokituta. Jotzen da GOP puntuazioen batez bestekoa eginez, murriztu egingo direla fonema-ezagutzaileko erroreak.

2. **Giza epaileengandik ikastea**: Ebakera balioesteko sistema automatiko batentzat, zentzuzko helburua da giza epaile baten moduan jokatzea. Giza jokaerara hurbiltzeko modu bat da gizakien etiketatzeko modutik ikastea, [134]n azaltzen den bezala: s hizlariak esandako q_i fonema batentzat entrenamenduko datu-base bateko agerpenetan giza epaile batek jarritako *oker* etiketen kopuru osoa $c_s(q_i)$ izanik, hizlari guztien oker etiketen kontaketa normalizatuen batezbesteko gisa zehaztu daiteke atalasea (ikus(6.4) ekuazioa).

$$T_{q_i} = log\frac{1}{S}\sum_{s=1}^{S}\left(\frac{c_s(q_i)}{\sum_{j=1}^{N} c_s(j)}\right) \tag{6.4}$$

   non N fonema kopuru osoa den eta S, entrenamendu-datuetako erabiltzaileen kopuru osoa. Normalizazioaren ondorioz, kontaketen batezbestekoak 0 eta 1 artean daude; hortaz, ateratzen diren balioen logaritmoak (6.3) ekuazioan zehazten direnen antzeko eskalako atalaseak dira.

3. **Errore-modelatze esplizitua**: Ebakera-akatsak bi akats mota nagusitan sailkatu daitezke. Lehen motan, banakako okerreko ebakerak daude, ikasle batek hitz baten ebakera ezagutzen ez duenean gertatzen direnak. Bigarren motan, xede-hizkuntzako soinuen ordezkapenak daude, ikaslearen jatorrizko hizkuntzan existitzen ez diren soinuenak. Bigarren motako akatsei *okerreko ebakera sistematikoak* deritze.

   GOP metodoak ez ditu erabiltzen ikasle guztien jatorrizko hizkuntza guztietako fonemen ereduak; beraz, okerreko ebakera sistematikoen kasuan, ez-jatorrizko hizketa ez da behar bezala modelatuko akustikoki. Hala ere, halako akatsen

detekzioa hobetu egin daiteke, GOP puntuazioan ikaslearen jatorrizko hizkuntzari buruzko ezagutza sartzeko aukera bagenu. Horretarako, ezagutza-sare bat inplementatu daiteke, bai ebakera zuzenak, bai ohiko ebakera okerrak erantsiz, fonema bakoitzerako akatsen azpi-sare gisa, xede-hizkuntzako eta iturburu-hizkuntzako fonema-ereduen multzoak erabiliz. Metodo horrek badu abantaila bat: iturburu-hizkuntza bakoitzerako eredu akustikoak lortu beharra.

4. **Akats simulatu edo artifizialak**: Fonema-atalaseak lortzeko beste modu bat da datu artifizialak erabiltzea. [46]n azaltzen denez, datu artifizialak ebakera-hiztegia manipulatuz sortzen direnak dira, alegia, ebakerak fonema desberdinak izateko aldatuz. Adibidez, /*aa*/ soinuaren agerpen guztiak /*iy*/ soinuez ordezkatu litezke, eta abar. Hala, leku jakinetan ebakera-akatsak dituzten hizketa-datuak sor daitezke. Ikerketa hartako esperimentuek azaltzen dute monofonema-ereduek emaitza hobeak ematen dituztela trifonema-ereduek baino, % 90eko puntuatze-zehaztasuna (SA, *Scoring Accuracy*) lortuz oker onartuen (FA, *False Acceptance*) % 8ko ratioaz, atalase optimorako. Ezarpen horietarako, ebaluazio-tresna bideragarria dirudi GOP puntuazioaren metodoak.

Euskararako, ez dago OBEL datu-baserik eskuragarri, eta, hortaz, ezin dugu ebakidura okerreko fonemen errealizaziorik lortu. Horren ondorioz, akats artifizialak (akats simulatuak) sortzeko konponbidera jo dugu. Hala ere, hiztegiko fonemak talde akustiko bereko fonemekin soilik ordezkatu dira; izan ere, fonema bat ebakitzen saiatzen garenean, probableagoa da "antzeko" soinu batez ordezkatzea (bokal bat beste bokal batekin, adibidez), "oso soinu desberdin" batez baino (bokal bat frikari ahoskabe batekin, adibidez).

Fonema taldeak

Fonema talde akustikoki antzekoak lortzeko, laborategian lehenago garatutako multzokatzelan bat erabili zen. [135]n azaltzen denez, nahaste bateko GMM ereduak entrenatu ziren fonema bakoitzeko, *Basque Speechdat* datu-basea erabiliz [97]. Erabilitako fonema-inbentarioa euskarazko SAMPA kodeko fonema multzoari dagokio[1]. Entrenamendua gauzatu zenean, erregresio-zuhaitzak baliatu ziren GMM ereduen parametro akustikoekin, talde desberdinak erauzteko. 6.1. irudian, lortutako fonema taldeen dendrograma ageri da.

Lan hartarako, ebakitze-puntu egoki bat hautatu zen (lerro horizontal marraduna), eta haren emaitza zortzi fonema talde izan ziren (bokalak talde bereko elementutzat jota, zeren, zentzu zorrotzean, bokal bakoitzak talde bat osatu beharko bailuke). Nabarmentzekoa da ateratako taldeek ia erabat bat egiten dutela euskarazko SAMPA kodeko fonemen artikulazio-moduekin (ikus 3.2.2. atala). Kontuan izan fonema herskari ahostunen alofono hurbilkariak diren B, D eta G fonemak ere bazirela sailkapenean, baina lan honetan saihetsi egin dira, eredu bakarra sortu baitzen bi aldaera fonetikoetarako;

1 http://aholab.ehu.eus/sampa_basque.htm

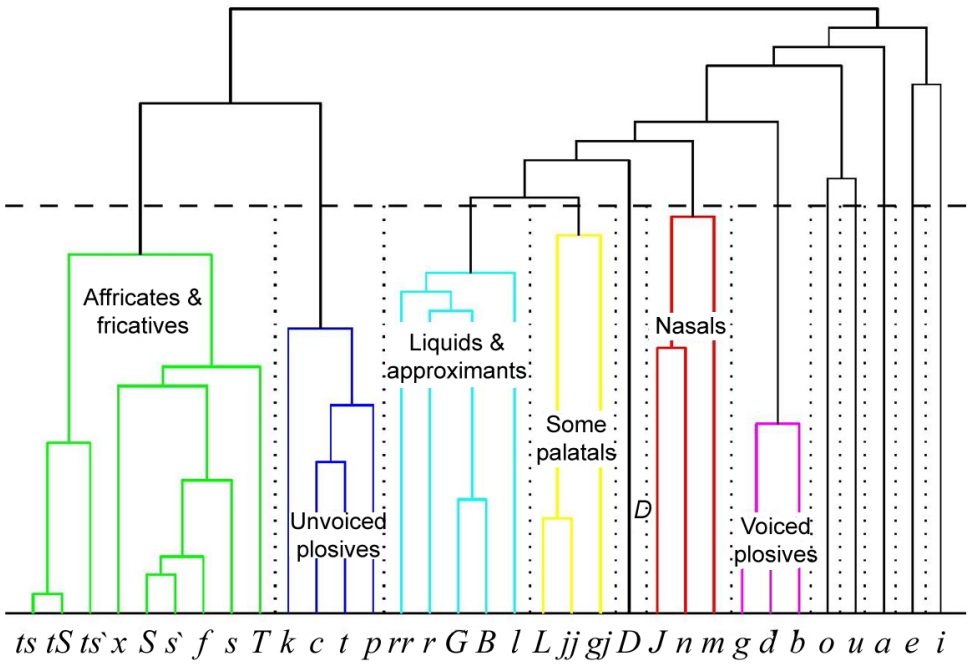


Figure 6.1: Fonema taldekatzearen irteerako dendrograma. Fonema multzo bakoitza kolore desberdin batez adierazirik dago.

geroago, erabaki horren zenbait ondorio ikusiko ditugu. Horrenbestez, zazpi talde geratu ziren. Bestalde, afrikatuen eta frikarien taldea bitan banatu zen, euskarak ezaugarri bereziak baititu bi talde horiei dagokienez, eta garrantzitsua izango da haiek lantzea. Hortaz, kontsideratu zen garrantzitsua dela talde bateko fonemak eta beste batekoak behar bezala bereizteko eredu gehiago entrenatzea.

Azkenik, zortzi talde zehaztu ziren:

- Bokalak: a, e, i, o, u
- Herskari ahoskabeak: c, p, t, k
- Likidoak: r, rr, l
- Afrikatuak: ts‘, ts, tS
- Sudurkariak: m, n, J
- Sabaikariak: L, jj, gj
- Herskari ahostunak: b, d, g
- Frikariak: f, x, T, s‘, s, S

Fonema taldeak zehaztuta, oker ebakitako fonemen GOP puntuazioak kalkulatu ziren, hiztegiko fonema bakoitza talde bereko beste batez ordezkatuz. HMMak testuinguruaren araberakoak direnez (trifonemak), aldatu egin zen fonema ordezkatuen testuingurumendekotasuna ere, hala transkripzioetako koherentziari eutsiz.

Basque Speecon-like datu-baseko *train* blokeko fitxategi guztiak erabiliz lortu ziren GOP puntuazioak, lerrokatze behartu moduan, eta prozesua hainbat aldiz errepikatu zen akats-simulazio desberdinekin, datu gehiago lortzearren. Orduan, bi banaketen histogramak kalkulatu ziren fonema bakoitzerako, fonema bakoitzaren puntuazio guztiak elkartuz. Hala, egiaztapen-atalaseak lortu ziren, GOP puntuazioen banaketa-funtzioetatik erauzitako errore berdinaren tasak (EER, *Equal Error Rate*) kalkulatuz, fonema bakoitzerako. Adibide gisa, /*a*/ fonemari dagozkion puntuazioen banaketa ageri da 6.2. irudian, bi dentsitate-funtzioekin.

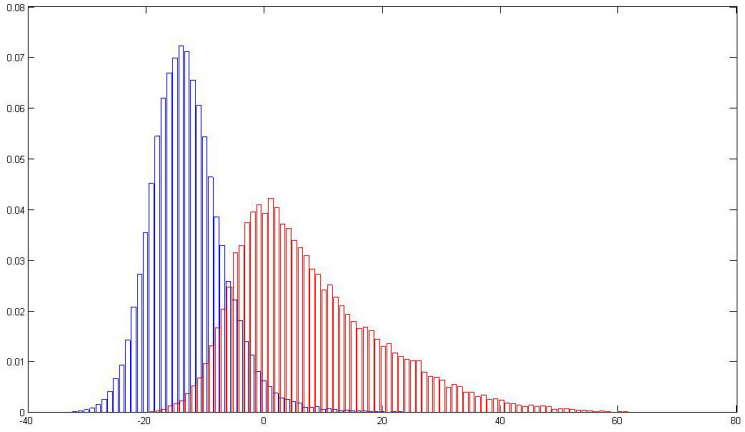


Figure 6.2: /*a*/ fonemaren GOP puntuazioen histograma normalizatuak: barra urdinek zuzen ebakitako fonemen GOP banaketak adierazten dituzte; barra gorriek, berriz, oker ebakitako fonemen GOP banaketak (akats simulatuetatik atereak).

Lehen begiratuan, ikusi zen bazirela fonema batzuk arazoak emango zituztenak, oso GOP banaketa gainjarriak zituzten eta. Datu-basean jatorrizko eta ez-jatorrizko euskal hiztunek fonema batzuk ebakitzean dituzten desberdintasunengatik gertatzen da hori; batez ere hizlarien ama-hizkuntzan existitzen ez diren fonemetarako, hala nola /*ts*‘/ edo /*s*‘/. HMMak entrenatzeko unean hori ez baitzen kontuan hartu, HMM berriak eraiki ziren, Euskal Herriko ekialdean jaio eta bizi diren jatorrizko hizlariei dagozkien seinaleak soilik erabiliz; izan ere, mendebaldean, gaur egun, /*s*‘/ fonemarik ez dago —/*s*/ fonemarekin bat egin zuen—. Hala, 155 hiztunetatik (*train* blokekoetatik) ekialdeko 76 euskal hiztunak hautatu ziren, HMM berriak entrenatzeko. HMM berri horiekin, prozedura osoa errepikatu zen, eta GOP banaketa berriak lortu ziren. Fonemarik arazotsuenen banaketen artean bereizte nabarmena lortu zen hala. Horren adibide bat 6.3. irudian ageri da, non jatorrizko hiztunen seinaleekin entrenatutako HMMak erabiliz GOP banaketa bereiziagoak lortzen baitira.

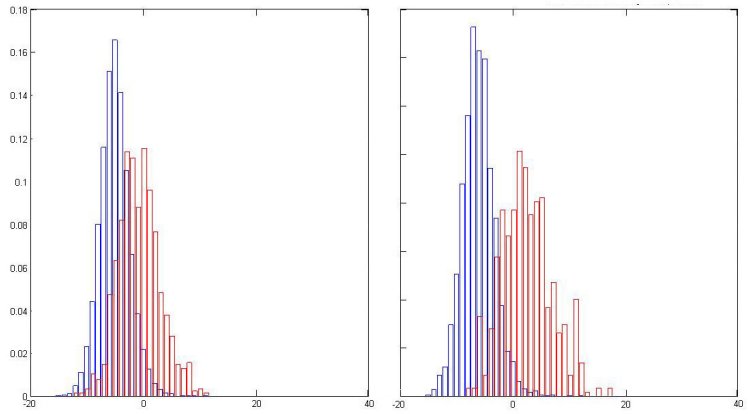


Figure 6.3: *ts‘* fonemaren GOP puntuazioen histograma normalizatuak, HMMak entrenatzeko hizlari guztiak erabiliz (ezkerrean) eta jatorrizko hiztunak soilik erabiliz (eskuinean).

Hala eta guztiz ere, badira oraindik ere fonema arazotsuak; esate baterako, /*ts*/ fonema, mendebaldeko euskalkietan /*ts‘*/ gisa ahoskatzen dena eta erdialdeko euskalkietan /*tS*/ fonematik hurbilago dagoena. Banaketa pareak ez ziren banatu HMM berriak erabiliz, 6.4. irudian ageri denez.

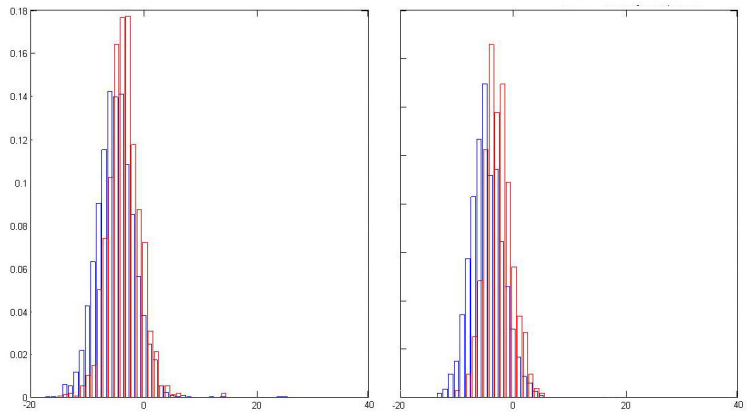


Figure 6.4: *ts* fonemaren GOP puntuazioen histograma normalizatuak, HMMak entrenatzeko hizlari guztiak erabiliz (ezkerrean) eta jatorrizko hiztunak soilik erabiliz (eskuinean).

HMMak entrenatzeko jatorrizko hiztunen audio seinaleak soilik erabiliz lortutako EERak 6.1. taulan ageri dira. Fonema batzuek EER balio baxua badute ere, beste batzuek balio altuagoak dituzte, eta hobetu egin beharko litzateke hori. Fonema herskari ahostunen kasuan, baliteke, arestian azaldu denez, bi errealizazio desberdin fonema berean batzeak eragin izana. Txistukarien kasuan, bi arazo sortzen dira: batetik, /*s‘*/

fonema /s/ gisa ahoskatzen da Euskal Herriko zenbait eremutan; bestetik, /ts/ fonema /ts‘/ gisa ahoskatzen da Euskal Herriko eremu batzuetan, eta /tS/ gisa beste batzuetan. Hortaz, soinu horiei dagozkien HMMak ez dira behar bezala entrenatu. Kontuan izan, bestalde, *gj* fonemak ez duela emaitzarik, fonema horren agerpenak oso urriak baitira datu-basean.

Table 6.1: Fonema bakoitzerako EER balioak, akats simulatuen metodoa erabiliz kalkulatuak.

Phone	EER	Phone	EER	Phone	EER	Phone	EER	Phone	EER
/a/	14.03	/p/	18.30	/ts‘/	10.61	/L/	19.14	/f/	12.85
/e/	18.64	/t/	20.54	/ts/	36.74	/jj/	32.05	/x/	5.10
/i/	07.99	/k/	22.34	/tS/	27.85	/gj/	-	/T/	24.62
/o/	15.53	/r/	17.11	/m/	15.66	/b/	30.09	/s‘/	34.82
/u/	10.92	/rr/	14.66	/n/	38.45	/d/	24.88	/s/	16.94
/c/	26.18	/l/	17.66	/J/	13.70	/g/	28.93	/S/	20.47

Atal honetan azaldutako lanaren zati handi bat nire ikerketa-egonaldian zehar egin zen 2012an Alemanian, Dresdengo Unibertsitate Teknikoan, Akustika eta Hizketa Komunikazioaren Institutuan. Xehetasun gehiagorako, ikus lan haren emaitza gisa argitaratutako artikulua: [67].

6.3 Lehendabiziko ebaluazioa

Emaitzak ebaluatzeko hautatutako neurria oso erabilia den puntuatze-zehaztasun (*SA*, *Scoring Accuracy*) koefizientea da, (6.5) ekuazioan ageri den moduan kalkulatzen dena.

$$SA\,(\%) = \left(\frac{CA + CR}{CA + CR + FA + FR}\right) \cdot 100 \tag{6.5}$$

non CA: zuzen onartutako elementuak (*Correctly Accepted*); CR: zuzen baztertutako elementuak (*Correctly Rejected*); FA: oker onartutako elementuak (*Falsely Accepted*); FR: oker baztertutako elementuak (*Falsely Rejected*).

6.3.1 Fonema-mailako testak

Diseinatutako estrategia balioesteko, bi esperimentu egin ziren, grabazio multzo desberdinak erabiliz. Bi testetan, ekialdeko jatorrizko hiztunekin entrenatutako HMM multzoa erabili zen. Fonema bakoitzaren GOP puntuazioa lerrokatze behartuaren prozedura erabiliz kalkulatu zen, eta EER atalasearekin alderatuz erabaki zen emaitza.

- **1. testa**: Test hori egiteko, Euskal Herriko ekialdean jaio eta bizi diren jatorrizko hizlariak hautatu ziren datu-basearen *test* bloketik (13 hizlari, 75etik, ikus 3.1.4. atala). Hizlari horiei dagozkien transkripzioak ez ziren ikuskatu, baina ontzat eman genuen ez zegoela akatsik eta hizlari guztiek behar bezala ahoskatzen zituztela fonema guztiak. Hortaz, suposatu dugu sistemak zuzentzat jotako fonemak zuzen onartuak (CA) direla, eta okertzat jotako fonemak, aldiz, oker baztertuak (FR).
- **2. testa**: Hori egiteko, euskaraz "behe-mailako" trebetasunak dituzten hizlarien grabazioak erabili ziren, guztira 25, 75etik. Nahiz eta aurrez ezin den jakin fonema bat zuzen ala oker ebaki den, batere ezagutze linguistikorik eta etiketarik gabe egin zen testa, 1. testean erabilitako fitxategi multzo berberarekin, begiratu batean ikusteko ea nolakoa izan litekeen sistemaren funtzionamendua "behe-mailako" ikasleekin.

6.2. taulan, lau fonema adierazgarriren emaitzak ageri dira: bi bokal (/*a*/ eta /*u*/), zeinak emaitza onak baitituzte; eta bi txistukari (/*ts*‘/ eta /*s*‘/), zeinak espainieran ez baitiren existitzen. /*a*/ eta /*u*/ fonemen SAek aldaketa txikiak dituzte test batetik bestera; horrek esan nahi du jatorrizko eta ez-jatorrizko hizlariek antzera ahoskatzen dituztela. /*ts*‘/ eta /*s*‘/ fonemen SAk, ordea, nabarmen jaisten dira, batez ere /*ts*‘/ren kasuan; horrek esan nahi du fonema horiek nahiko desberdin ahoskaturik daudela.

Table 6.2: /*a*/, /*u*/, /*ts*‘/ eta /*s*‘/ fonemen errealizazio kopuruak eta SAk, 1. eta 2. testetan.

		/*a*/	/*u*/	/*ts*‘/	/*s*‘/
1. testa	#Errealiz.	5 524	1 937	750	1 317
	SA (%)	86.22	89.67	83.73	74.26
2. testa	#Errealiz.	9 923	3 481	1 438	2 469
	SA (%)	84.06	87.33	41.59	49.49

Azkenik, beste esperimentu bat egin zen:

- **3. testa**: 2. testean SArik txarrenak izan zituzten fonemekin egindako esperimentua. "Behe-mailako" $L2$ trebetasunen azpicorpuseko fonemen errealizazioak zuzen ala oker ebakitakotzat etiketatu ziren, eskuz. /*ts*‘/ fonemaren kasuan, 813 errealizazio zeuden; horietatik 375 zuzen eta 438 oker gisa etiketatuak. /*s*‘/ fonemarentzat, 1 348 errealizazio etiketatu ziren; 720 zuzen eta 628 oker gisa etiketatuak. Etiketa horiek kontuan izanda, GOPak kalkulatu ziren berriro, eta 6.3. taulako emaitzak lortu ziren.

Ikus dezakegu ezen 3. testeko SAk (6.3. taula) 1. testekoetatik gertuago daudela orain (6.2. taula), 2. testeko emaitzekin alderatuta. Horrek baieztatzen du datu-basean

Table 6.3: Automatikoki sortutako etiketak eta eskuz esleitutako etiketen konparazioen emaitzak, *ts‘* eta *s‘* fonementzat.

3. testa	/*ts‘*/	/*s‘*/
CA (%)	32.84	33.97
CR (%)	43.67	29.15
FA (%)	10.21	17.43
FR (%)	13.28	19.43
SA (%)	76.51	63.13

badirela oker ebakitako fonemak. Adibidez, /*ts‘*/ fonemaren emaitzek erakusten dute HMMak jatorrizko hiztunen datuekin soilik entrenatuz lortzen den banaketa-bereizketa (ikus 6.3. irudia) erabilgarria izan dela erabaki-atalaseak kalkulatzeko. Bestalde, /*s‘*/ fonemaren emaitza ez hain zehatzak interpretatzeko, kontuan hartu behar da gaur egun /*s‘*/ soinua ez dela existitzen Euskal Herriko eremu handi batean, eta ekialdeko zenbait eremutako jatorrizko hiztunek ere jasaten dute horren eragina.

Ikuspuntu orokor batetik, emaitzek erakusten dute gure *AhoSR_L2* sistema gai dela esakuntza batean bai ebakera zuzeneko fonemak, bai ebakera okerreko fonemak hautemateko, batez ere ikasleen $L1$en existitzen ez diren fonemetan arreta jarrita. Ikusi dugu alde handiak daudela "behe-mailako" azpicorpuseko ikasleen trebetasunen artean, seguruenik datu-basearen sailkapenak bi trebetasun-mailaren artean soilik hautatzeko aukera ematen duelako: behe-maila eta goi-maila. Sistemak, espero bezala, hobeto bereizten ditu behe-mailako hizlarien okerreko ebakerak.

Emaitzak [125]en lortutakoekin alderatuz, SA baxuagoak lortu ditugu hizlariaren $L1$ean existitzen ez diren fonementzat; baina emaitza hobeak hizlariak bere jatorrizko hizkuntzan dagoeneko badituen fonemetarako. Hasiera-puntu gisa, esan dezakegu ezen erabakia hartzeko atalaseak kalkulatzeko estrategia baliozkotu egiten dutela lortutako emaitzek.

Lan horri buruzko xehetasun gehiagorako, ikus [136] eta [137].

6.3.2 Hitz-mailako testak

Hitz-mailako puntuazioak ere aztertu ziren hasierako test haietan. Asmoa zen ikustea ea hitz-mailako puntuazioak nahikoa —edo, gutxienez, erabilgarri— diren hitzez hitzeko esaldi-egiaztapenerako (HHEE), edo fonema-mailako puntuazioak ere kontsideratu beharra dagoen.

Hitz-mailan, erraz zehaztu daiteke hitz baten fonema-puntuazio (PS, *phone score*) osoa: hitza osatzen duten fonemen GOPen batura haztatua, (6.6) ekuazioan adierazten

den bezala.

$$PS\,(hitza) = \sum_{u=1}^{N} w_u \cdot GOP\,(y_u) \tag{6.6}$$

non w_u: hitza osatzen duten N fonemen artean u posizioan dagoenaren haztapena den. Eskuarki, haztapenak berdinak izan ohi dira fonema guztientzat [138].

Hitz-mailako atalaseak fonementzat erabilitako metodologia bera erabiliz kalkulatu ziren (akats simulatuak erabiliz oker ebakitako hitzen GOPak lortu eta EER bat kalkulatuta). Horrenbestez, hiru test egin ziren, hala kalkulatutako atalaseen baliozkotasuna aztertzeko:

- **1. testa**: Sistema *Basque Speecon-like* datu-baseko *test* ataleko 2 218 esaldirekin testatzean datza, 7 296 hitz guztira. Esperimentuan, transkripzio guztiak zuzenak dira; hortaz, sistemak zuzen gisa etiketatutako hitzak CA dira, eta oker gisa etiketatuak FR dira.
- **2. testa**: *Basque Speecon-like* datu-baseko *test* ataleko 1 174 hitz isolatu testatzean datza, fitxategi bakoitza hiztegiko ausazko hitz batekin, zuzena ez den batekin. Hala, esperimentu horretan zuzen gisa etiketatutako hitzak FAk dira; oker gisa etiketatutakoak, berriz, CRak.
- **Test 3**: Zuzen eta oker ahoskatutako hitzak dituen hizketa testatzean datza. Horretarako, *Basque Speecon-like* datu-baseko *test* ataleko 886 esaldi erabili ziren, transkripzio-fitxategietako esaldietan hitz bana ezabatuta; hala, okerki erantsitako hitz bat simulatu zen esaldi bakoitzean. Jatorrizko sarrerako transkripzio fitxategiek 5 080 hitz dituztenez, 886 hitz okerrak lirateke; gainerako 4 194ak, zuzenak.

Hiru esperimentuetan lortutako emaitzak 6.4. taulan ageri dira. Nahiz eta aurrerago kalkulatutako puntuazio-banaketen emaitzak onak izan, 3. testean CAren estaldura % 99.12 da, eta CRaren estaldura % 84.88 (CAren estaldura: zuzen onartutako hitzak hitz

Table 6.4: Hitz-mailako lehendabiziko 1., 2. eta 3. testen emaitzak.

	1. testa	2. testa	3. testa
CA	7 090	—	4 157
CR	—	1 174	752
FA	—	0	134
FR	206	—	37
SA (%)	97.18	100.00	96.63

zuzen guztien artean; *CR*ren estaldura: zuzen baztertutako hitzak, hitz oker guztien artean). Horrek esan nahi du hitz okerrak ez direla zuzenak bezain ondo klasifikatzen; espero zenaren kontrako asimetria bat EERak erabiliz atalaseak kalkulatzean.

Lehendabiziko esperimentu horiek agerian uzten dute sistema ingurune errealistago batean ebaluatu beharra, ebaluazio fidagarriago bat lortzeko. Lehendabiziko esperimentu horiei buruzko informazio gehiagorako, ikus [95].

6.4 Softwarea

HHEE esperimentuak ingurune erreal batean egiteko, erabiltzailearen interfaze grafiko bat (GUI, *Graphic User Interface*) eraiki zen *AhoSR*n (ikus 6.5. irudia). GUIa C++ lengoaian diseinatu zen *Windows*erako, kontuan izanda batezbesteko erabiltzaileak ohituago daudela *Windows* sistema erabiltzen beste edozein sistema eragile baino.

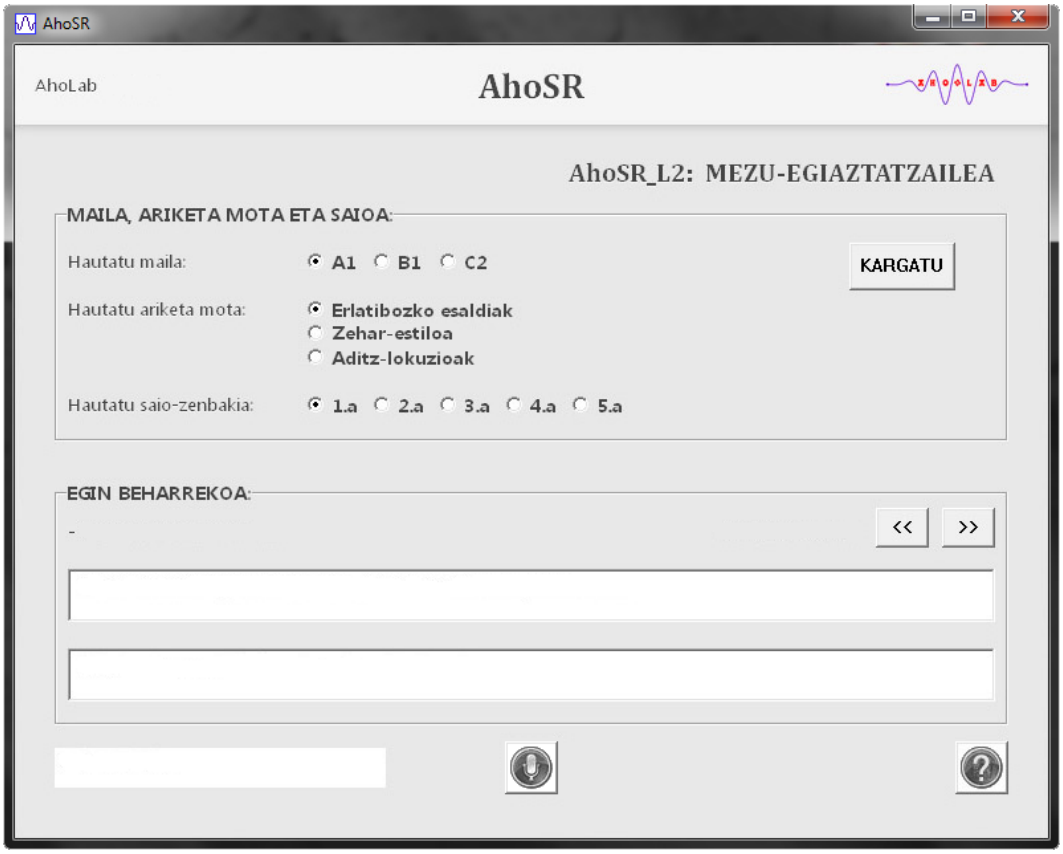


Figure 6.5: AhoSR_L2 sistema.

GUIak bi bloke nagusi ditu: lehen blokean, erabiltzaileak gaitasun-maila desberdinetariko bat hauta dezake (A1, B1 eta C2; ikus [139]), eta gramatika-ariketa mota nahiz

ariketa-zenbakia hautatzeko aukera du. Bigarren blokean, ariketa aurkezten da, baita mikrofono-botoi bat ere ariketa ebazten hasteko. Erabiltzaileak klikatutakoan, botoia gorri bihurtzen da adieraziz deskodetzailea mikrofonotik audioa hartzen hasi dela. Sistemak, orduan, erabiltzaileak esatea espero den hitz bat egiaztatzen badu, kaxa batean erakutsiko da; hala, erabiltzaileak berehalakoan jakin dezake ondo ari den ariketa ebazten. 6.6. irudian, une horren adibide bat ageri da, non sistemak zenbait hitz egiaztatu eta azpiko kaxan pantailaratu baititu. Erabiltzailea ariketaren amaierara iristen bada edo aurrez zehaztutako denbora-tartea (20 s) agortzen bada, audio-sarrera eten egiten da eta mikrofono-botoia itzali egiten da berriro. Erabiltzailea ez bada ariketaren amaierara iritsi, "?" botoia erabil daiteke, falta den emaitzaren zatia bistaratzeko.

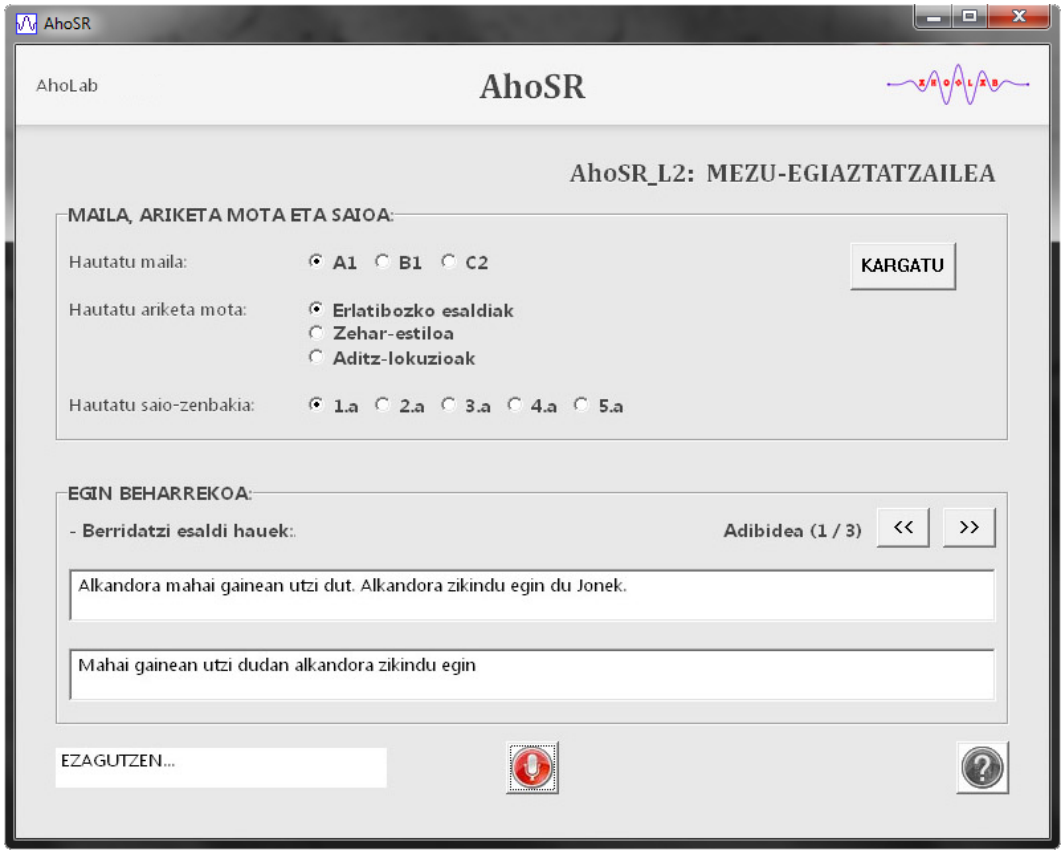


Figure 6.6: AhoSR_L2 sistema, abian.

Komeni da kontuan izatea ezen, esandako hitz bat pantailan agertzen ez bada, balitekeela erabiltzaileak okerreko hitz bat esan izana (CR) edo hitza ez izana behar bezala egiaztatu (FR).

6.5 Ariketak eta ebaluazioaren diseinua

Sistema $L2$ eskuratzeko benetako ingurune batean ebaluatzeko, gramatika- eta sintaxi-ariketa multzo bat hautatu zen, hasiera batean euskarazko A2 gaitasun-mailarako (oinarrizko maila). Ariketa mota desberdinak pentsatu ziren, hiru euskara-irakasleren laguntzarekin. Ariketak kontu handiz diseinatu ziren, hitz-ordena zorrotza behar duten gramatika-kontzeptuak lantzeko, hala nola erlatibozko esaldiak osatzea, zeharkako estiloa edo aditz-esapideak. 300 esaldi desberdin erabili ziren, guztira, ebaluazioa egiteko.

Tresnaren ebaluazioa euskara irakasten duten bi erakundetan egin zen. Guztira, 20 boluntariok (10 gizonezko eta 10 emakumezko) hartu zuren parte atazan. Denak ziren A2 mailakoak, eta ordurako ikasgelan ikasia zuten probako materiala. Boluntario guztien ama-hizkuntza espainiera zen, batentzat izan ezik, harentzat katalana baitzen. 6 ikasle 20-29 urte bitartekoak ziren, 12 ikasle 30-39 urte bitartekoak, eta 2 ikasle 40-49 bitartekoak. Boluntarioen ezaugarri nagusiak 6.5. taulan ageri dira.

Table 6.5: Parte hartutako ikasleen ezaugarriak.

Characteristic		
Gender	Gizonezkoak	10
	Emakumezkoak	10
Age	20-29	6
	30-39	12
	40-49	2
$L1$	Espainiera	19
	Katalana	1

Ikasle bakoitzak 30 ariketa ebatzi behar zituen, 3 bloketan banatuak (bloke bakoitza ariketa mota desberdin bati zegokion). Ikasleei, hasi aurretik, 3 adibiderekin trebatzeko aukera ematen zitzaien. 3 adibideak ebatzi bitartean, sistemari edo ariketa motari buruzko edozein galdera egin zezaketen, baina gero ez. Testak ordenagailu berean egin zituzten 20 boluntarioek. Gainera, USB bidezko kasko-entzungailu mikrofonodun berbera erabili zuten denek.

Testaren ondoren, ikasleei galdetegi labur bat eman zitzaien, haien esperientziari buruzko *feedback* orokor bat lortzearren. Galderak 0 eta 10 bitartean puntuatzekoak ziren, eta puntu hauei buruzkoak ziren:

- 1. galdera: Sistemaren funtzionamendua.
- 2. galdera: Sistemaren erabilgarritasuna.
- 3. galdera: Sistemaren erabiltzeko erraztasuna.
- 4. galdera: Sistema hizkuntz-eskoletan erabiltzeko gomendagarritasuna.
- 5. galdera: Balorazio orokorra.

6.6 Emaitzak

Ebaluazio-prozesuan ebatzitako ariketa kopurua 600 bazen ere, 597 audio-fitxategi gorde zituen tresnak; izan ere, 3 erabiltzailek ariketa bana saltatu egin zuten, nahi gabe. Testaren ondoren, fitxategi guztiak eskuz etiketatu ziren, eta esakuntza egiaztatzeko tresnaren bidez ebaluatu ziren. Lortutako puntuatze-zehaztasun osoa eta CA, CR, FA eta FR 6.6. taulan ageri dira. Guztira 2 952 hitz ebaluatu ziren, eta ikasleek esandako hitz kopurua 4 404 zen. Azken zenbaki hori ez da zehatz-mehatz CA, CR, FA eta FRren batura (4 402), espero zitekeen bezala, zeren eta, bi kasutan, (oker) egiaztatutako bi hitz hitz bakarrari dagokie.

Table 6.6: Ingurune erreal batean egindako HHEE esperimentuaren hitz-mailako puntuatze-zehaztasuna.

CA	2 419
CR	1 538
FA	136
FR	309
SA (%)	89.89

Ingurune errealean egindako ebaluazio horretan lortutako puntuatze-zehaztasuna txikiagoa da, espero bezala, lehenengo testetan lortutakoak baino (ikus 6.3. atala). Egiaztatzeko tresna bizitza errealean erabiltzen denean, erabiltzaileek akatsak egiten dituzte, eta bazterketak nabarmen handiagotzen dira laborategiko esperimentuekin alderatuz. Horrek agerian uzten du tresnaren erabilgarritasuna eta ingurune erreal bat erabili beharra halako aplikazioak ebaluatzeko unean.

CA eta CR-ren estaldura eta doitasuna 6.7. taulan ageri dira. Datuei dagokienez, ondorioztatzen dugu bi kasuetan sistemak antzera jokatzen duela, bi estaldura balioak konparagarriak baitira; doitasunari dagokionez, berriz, zeina uler baitaiteke onartutako hitz guztien artean zuzen onartutako hitzen ehuneko gisa (CA doitasuna) eta baztertutako hitz guztien artean zuzen baztertutako hitzen ehuneko gisa (CR doitasuna), ikus daiteke asimetria txiki bat dagoela bien artean; izan ere, sistemaren funtzionamendua hobea da onartutako hitzei dagokienez.

Table 6.7: Ingurune erreal batean egindako HHEE esperimentuko CAren eta CRaren estaldura eta doitasuna.

	CA	CR
Estaldura (%)	88.67	91.88
Doitasuna (%)	94.68	83.27

Sistemak egindako erroreak analizatuz (FA gehi FR), akatsak dituzten 235 fitxategi

daude, 597tik. Fitxategiko (edo esaldiko) errore kopuruen banaketa 6.7. irudian ikus daiteke. Irudiak ikus daitekeenez, fitxategiko gehienezko FA kopurua 2 da; gehienezko FR kopurua, berriz, 9. Testean zehar, ikasleek hitz zuzen bat esan baina sistema, ordea, egiaztatzeko eta pantailaratzeko gai ez zenean, ikasleek behin eta berriro errepikatzen zuten hitz hori. Hori, batez ere, hitz laburrekin gertatzen dela ikusi dugu, nahiz eta konfiantza-puntuazioari bilbe kopurua normalizatzeko faktore bat aplikatu. Ohartu ginen ezen fonema jakin batzuk —hurbilkariek, adibidez— sarriago agertzen zirela akatsen artean. Gainera, hauteman genuen ezen, FR akatsen eraginez, sistema ez zela gai leheneratzeko, fitxategi guztien % 17.45 kasutan.

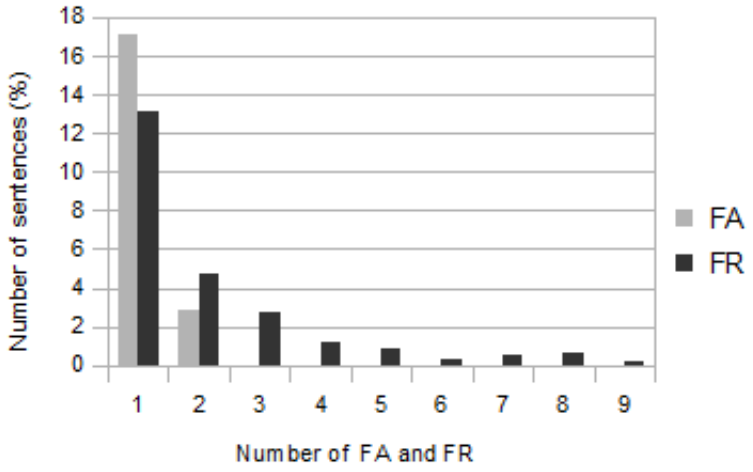


Figure 6.7: FA eta FR kopuruen banaketak fitxategien artean.

Erabiltzaileek betetako galdetegiari dagokionez, emaitzak 6.8. taulaan ageri dira. Erabiltzeko erraztasuna da altuen puntuatutakoa. Tresnaren balorazio orokorrak 8.28ko puntuazio lortu badu ere, funtzionamenduak 7 lortu du, puntuaziorik baxuena. Horrek esan nahi du ezen, sistema gomendagarria bada ere (% 90 inguruko puntuazioa), baxuagoa dela erabiltzaileek haren funtzionamenduari buruz duten pertzepzioa. Bistakoa da erabiltzaileentzat ez dutela garrantzi bera sistemaren akatsek eta asmatzeek, zeren, berez, sistemak ondo funtzionatzea espero baitute.

Table 6.8: Ikasleen galdetegian lortutako batez besteko puntuazioak.

Galdera zenbakia	1	2	3	4	5
Batez besteko puntuazioa	7.00	8.25	9.15	8.95	8.28

6.7 Konklusioak

Kapitulu honetan, OBEL eta HHEE sistemen lehenengo esperimentuak deskribatu dira. OBEL sistemaren helburua da esandako soinu bakoitzarentzat puntuazio bat lortzea. HHEE sistema gramatika-ariketak (AGP) egin ahala ebazteko metodo bat da; hala, erabiltzaileek berehalakoan lortzen dute esandako edukiaren *feedback*a. Horrek aukera ematen die erabiltzaileei beren ustea aldatzeko, esandako hitz ordena —adibidez— zuzena ez bada.

GOP atalaseak ezartzeko, zuzen eta oker ahoskatutako fonemen GOP banaketak behar dira. Hala ere, ez dago egiaztapenerako datu-baserik euskararako, eta ASR estandarrek ez dute oker ebakitako daturik. Hala, oker ebakitako fonemen GOPa kalkulatzeko metodo bat asmatu da: artifizialki sortutako ebakera-akatsak. Metodo horren funtsa da aldaketa isolatuak (edo akats simulatuak) txertatzea hiztegian, fonema jakin batzuk beste batzuekin ordezkatuz; emaitza kontserbadoreagoak lortzeko, talde akustiko bereko fonemez ordezkatuz egin dira fonema-aldaketak. Fonema bakoitzerako entrenatutako GMMak taldekatuz lortu dira talde horiek, eta euskararako SAMPAren talde fonetiko ia berdinak dira.

Zuzen eta oker ahoskatutako fonemen banaketetatik lortutako EER erabiliz, atalaseak erauzi dira. Atalase horiekin, hainbat esperimentu egin dira fonema-mailan. Jatorrizko hiztunen datuak erabiliz, HMM eta GOP hobeak lortzen dira, zenbait fonema arazotsu desberdintzeko.

Hitz-mailako esperimentuak ere egin ziren, HHEE atazan zer jokaera duten ikusteko asmoz. Laborategiko esperimentuek oso emaitza onak eman zituzten (% 97.18, % 100.00 eta % 96.63ko *SA*, hiru testetan), baina sistemak baldintza errealistagoetan testatu beharra zuen. Hala, GUI bat eraiki zen sistemarako, eta esperimentu errealista bat egin zen A2 oinarrizko mailako 20 euskara ikaslerekin. % 90.00 inguruko *SA* lortu zen esperimentuetan (hitz-mailan), laborategian lortutakoak baino emaitza txarragoa, baina, hala ere, onak, ildo berean ikertzen eta sistema hobetzen jarraitzeko.

Dena dela, sistema horrek zenbait hobekuntza behar ditu inplementazio errealista baterako. Gaur egun, joera da hizketa-ezagutzailea zerbitzari batean instalatzea eta erabiltzaileak urrutitik konektatzea hara. Horrek esan nahi du erabiltzaile bakoitzak gailu desberdin baten bidez jasoko duela audioa; hortaz, behar-beharrezkoa da parametroen normalizazioa. Gainera, VAD bat ere beharrezkoa da halako sistema batean, audio-seinale jarraitu batetik hizketa-segmentuak erauzteko eta, prozesamendu-denbora irabaztearren, isilune-bilbeak baztertzeko.

Hurrengo atalean, *online* VAD teknika berri bat (ikus 8. kapitulua) eta *online* CMVN teknika berri bat (ikus 9. kapitulua) aurkeztuko ditugu, zeinak metodo berri hau dute oinarrian: normalizazio anitzeko puntuatzea (MNS, *Multi-Normalisation Scoring*). Gainera, erabakiak hartzeko prozesua hobetu egin da neurona-sare artifizialak erabiliz (ikus 10. kapitulua). Horren guztiaren emaitza webean oinarritutako OBEL eta AGP sistema da, sarbide unibertsalekoa.

PART III

Sistemaren hobekuntzak

CHAPTER 7

Online inplementazioa

7.1 Sarrera

Hemen aurkeztutako tresnaren helburua da sarbide unibertsala izatea, bezeroa edonon dagoela ere. Horretarako, ebazpiderik ezagunena da zerbitzari bat inplementatzea, bezeroak eskaera bat egiten duenean eragiten den prozesaketa guztia hantxe egin dadin. Zerbitzariak prozesaketa amaitu bezain laster, emaitza bat itzultzen zaio bezeroari.

Gaur egun, bezero-zerbitzari arkitekturako sistemarik hedatuena web-nabigatzailea da, baina duela gutxi arte, webeko audio-prozesaketa guztia oso oinarrizkoa izan da, eta *Flash*, *Java* appletak edo antzeko *plugin*ak erabili behar izaten ziren horretarako. HTML5eko *<audio>* elementuak aukera eman du audio *streaming*ak erreproduzitzeko, eta, oraindik ere garatzen ari den arren, etorkizunean aukera sofistikatuagoak izango ditu audioa nahasteko, prozesatzeko eta iragazteko funtzionaltasunak izango baititu, audioa prozesatzeko gaur egungo mahai gaineko aplikazio modernoetan aurki daitezkeen bezalakoak. Horrenbestez, badirudi web-nabigatzailea oso oinarri ona dela gure tesian aurkezten dugun moduko aplikazio-arkitekturak inplementatzeko. Hurrengo atalean (7.2. atala), xehetasun handiagoz deskribatzen da sistema-arkitektura *online* inplementaziorako.

Bestalde, CMVN teknika moldatu egin behar izan da. *Offline* kasuan, batezbestekoak eta bariantzak kalkulatzeko, prozesatzen ari den audio-segmentu osoko bilbe guztiak erabiltzen dira. *Online* kasuan, hasieran batezbestekoen eta bariantzen balioak estimatu behar dira eta egokitu gero, parametro cepstralek ahalik eta deformaziorik txikiena jasaten dutela ziurtatzeko. *Online*ko egokitzapen-ataza hori gauzatzeko estrategia desberdinei buruzko azterketa 9.3. atalean ageri da.

Azkenik, zenbait ondorio iruzkindu dira 7.3. atalean.

7.2 Web teknologia: HTML5

Gaur egun, HTML5 espezifikazioak, *Javascript*ekin batera, audio mikrofono batetik jasotzeko aukera ematen du Javascript bidez erabil daitekeen *Web Audio* APIaz [140] eta haren *MediaStream* interfazeaz [141]. Gaur egungo nabigatzaile-garatzaileak pixkanaka-

pixkanaka ari dira HTML5eko funtzionaltasunak eta APIak nabigatzaileei eransten (batek ere ez du oraindik HTML5eko betekizun guztiak % 100ean betetzen[1]), eta luze jo du *Web Audio* APIaren inplementazioa ere orokorra izan den arte. Dena dela, 2014tik nabigatzaile ezagun ia guztiek dute inplementaturik.

Web Audio APIaren paradigma nagusia audioa bideratzeko grafo bat da, non elkarri konektatutako hainbat *AudioNode* objektu dauden audioaren erreprodukzio orokorra zehazteko. Testuinguru horretan, *Mediastream* interfazeak audio- edo bideo-edukien fluxuak jasotzen ditu lokaleko multimedia gailuetatik, hala nola mikrofonoetatik edo bideo-kameretatik, eta zerbitzarira igotzeko prest uzten da, gero han prozesatu dadin.

Ordenagailu Bidezko Ebakera Lanketa (OBEL) atazan, grabatutako audioa ez da bidaltzen erabiltzaileak "bidali" botoia klikatzen duen arte. Hala, grabatutako datuak egiaztatu ditzake, bidali aurretik, erabiltzaileak. Hortaz, jotzen badu grabatutako esaldia ez dagoela behar bezala grabatuta edo akatsen bat daukala, audio-datuak baztertu egin daitezke eta berriro grabatu. Hitzez Hitzeko Esaldi Egiaztapena (HHEE) atazan, berriz, bestelako testuingurua behar da. Grabagailuko bilbe-blokeak iritsi ahala zerbitzarira bidali behar dira audio-datuak. Hala, erabiltzaileak hitz zuzena esan duen hauteman dezake zerbitzariak, eta, hala bada, *feedback*a bidali une horretan bertan, nabigatzailean audioa jasotzeko eta bidaltzeko prozesuak aurrera jarraitzen duen bitartean. Noranzko biko komunikazioa ezarri beharra dago, beraz.

7.2.1 OBEL sistemaren arkitektura

Ataza honetan, erabiltzaileak ez du espero denbora errealeko erantzunik. Sistemak esaldi bat irakurtzeko eskatuko dionez, askoz egokiagoa da datuei nabigatzailean eustea erabiltzaileak, grabatu eta gero, ondo dagoen egiaztatzeko aukera izan dezan. Hala, berriz grabatzeko aukera izango luke esaldia behar bezala ahoskaturik ez balego edo, adibidez, audio-seinalean kanpoko zarataren bat sartuko balitz. Erabiltzaileak grabazioaren kalitatea ona dela uste duenean, *BIDALI* botoia klikatuz bidal daiteke.

Audio-datuak *wav* fitxategi batean gorde ditzake nabigatzaileak. Hala, zerbitzariko *php* script batez, fitxategiaren burua irakur daiteke eta, hala, behar diren aldaketak egin. Lan honetako HMMak, esate baterako, 16 kHz-eko audio-seinaleez entrenaturik daude; beraz, eskuarki, txikiagotu egin beharko da laginketa-maiztasuna. Ondoren, *AhoSR* exekutatzen da *WAV* moduan, eraldatu berria den audio-fitxategia pasaraziz audio-sarrera gisa. Emaitza lortzen denean, nabigatzaileari itzultzen zaio AJAX (*Asynchronous JavaScript and XML*) erabiliz; horrek web-orriaren atal bat eguneratzeko aukera ematen du, orrialde osoa freskatu beharrik gabe.

Bestalde, audio-fitxategiak zerbitzari batean gorde daitezke, etorkizuneko ikerketa-lanetan erabiltzeko.

1 https://html5test.com/results/desktop.html

7.2.2 HHEE sistemaren arkitektura

Nabigatzailetik zerbitzarira audio-datuak lagin-blokeka bidaltzeko, HTML5en beste ezaugarri bat erabili dugu: *Web Sockets* teknologia [142]. Web aplikazioetarako noranzko biko komunikazioa ahalbidetzen duen teknologia da, socket bakarrean zehar funtzionatzen duena, eta HTML5ekin bateragarriak diren nabigatzaileetan *Javascript* interfaze baten bidez inplementatzen dena. Hala ere, *websocket*ak ez dira ohiko *socket*ak; praktikan, UDPren eta TCPren zenbait ezaugarri konbinatzen ditu: mezuetan oinarrituta dago, UDP bezala, baina fidagarria da, TCP bezala. *Websocket*ek aukera ematen diete nabigatzaileei zerbitzariekin pseudo-konexio bat zabaldu, datuak trukatu, eta, komunikazioa amaitu denean, ixteko. Bide egokia da audio-datuak modu antolatu eta eraginkor batean zerbitzarira bidaltzeko, zeren HTTP egoerarik gabeko protokoloa baita, hau da, ez du gordetzen aurreko konexioei buruzko informaziorik. *Websocket*ak ia nabigatzaile gehienek dituzte inplementaturik.

Bestalde, zerbitzariak,*websocket*ak erabiliz, konexio bat baino gehiago eta aldi berean kudeatu behar ditu. Zerbitzariaren aldeko web-aplikazioak garatzeko ingurune arin eta eraginkor bat *Node.js* da. Kode irekiko exekuzio-ingurunea da, plataformarekiko independentea, eta aplikazioei web zerbitzari gisa jarduteko aukera ematen dien berezko liburutegiak ditu. Gertaerek gidaturiko arkitektura du *Node.js*k, baita blokeorik gabeko I/O API bat, denbora errealeko web aplikazioetarako errendimendua eta eskalagarritasuna optimizatzeko diseinatua. Aplikazioak *JavaScript*en idatzita daude, eta, *Node.js* ingurunearen barnean, sistema eragile guztietan exekuta daiteke. Haren ostatatzearen eta mantentze-lanen arduraduna *Node.js Foundation*[1] da, *Linux Foundation*[2] erakundearen elkarlaneko proiektu bat.

Aurreko guztia kontuan izanda, HHEE sistemaren arkitektura hiru bloke desberdinez osaturik dago:

- Nabigatzailea: erabiltzaile-interfazea duen atala da. *Websocket* bezeroa dago bertan inplementatuta, eta konexio-eskaera bat egiten dio urrutiko *websocket* zerbitzariari (*Node.js* zerbitzarian). Konexioa ezarritakoan, nabigatzailean kokatutako *Javascript* programa baten bidez, mikrofonoko audioa jaso eta prozesatu egiten da (16 kHz-etara jaisten da laginketa-maiztasuna) datuak etorri ahala, eta zerbitzarira bidaltzen dira. Zerbitzaritik ere *feedback*a jaso daiteke, erantzunik egonez gero.
- *Node.js* zerbitzaria: Nabigatzailearen eta *AhoSR*-ren arteko bitartekaria da. *Websocket* zerbitzari gisa jokatzen du nabigatzailearekin, baita *websocket* bezero gisa ere, *AhoSR*-rekin konexio bat ezartzeko. Nabigatzailetik konexio-eskaera bat iristen denean, *Node.js* zerbitzariak *AhoSR* abiarazten du, eta *socket* konexio bat eskatzen du. Baldin eta, *socket*a ezartzean, errorerik gertatzen ez bada, *websocket*

1 https://nodejs.org/en/foundation/
2 http://collabprojects.linuxfoundation.org/

konexioa ere ezartzen da nabigatzailearekin. *AhoSR*tik *READY* mezua jasotzen duenean, nabigatzailera bideratzen du mezua, eta HTML bidezko grabagailua bistaratzen da han.

- *AhoSR*: *Socket* moduan exekutatzeko konfiguratu behar da, non audio-laginen blokeak *socket* konexio batean zehar iristen baitira. *Socket*a ezarrita baldin badago, lehen datu-blokea iritsi ahala ekiten dio hitza egiaztatzeko prozesuari, eta, hitz bati antzematen zaion bakoitzean, berehalakoan bidaltzen zaio *socket* bezeroari (*Node.js* zerbitzarian). Azken hitza detektatu denean edo denbora amaitu denean, *AhoSR*k *FINISH* mezua bidaltzen dio bezeroari, zeina nabigatzailera birbideratzen baita audioa jasotzeko prozesua gerarazteko.

Hiru blokeen arteko komunikazio-protokoloa 7.1. irudian ageri da. Kontuan izan testu berdea *websocket* bezeroaren eta zerbitzariaren arteko komunikazioari dagokiola; testu urdina, berriz, *socket* bezeroaren eta zerbitzariaren komunikazioari; eta testu gorriak erabiltzailearen ekintzak adierazten ditu.

Ordenagailu bakoitzak laginketa-maiztasun desberdin batean grabatzen duenez, audio-laginak prozesatu egin behar dira. Orain arte, *Mediastream* interfazeak ez du uzten laginketa-maiztasuna aldatzen; hortaz, berariazko funtzio bat gehitu zaio *audio context*ari *Javascript*en. Nabigatzaileak audio-seinaleen laginketa-maiztasuna 16 kHz-etara jaitsi eta zuzenean bidaltzen ditu *AhoSR*-ra, *Node.js* zerbitzarian prozesaketa gehiagorik gabe. Maiztasun-egokitzapen hori egin ezean, *AhoSR*k ezingo luke seinalearen laginketa-maiztasuna zein den asmatu.

Sistemak behar bezala funtzionatzeko faktore erabakigarri bat da audio-laginak uzteko datu-bufferraren tamaina. Datu-bufferra beteta dagoenean, gertaera bati deitzen zaio, eta une horretan bloke osoa bidaltzen da. Horrek esan nahi du sistemaren bereizmenari eragingo diola horrek. Tamainak hauetariko bat izan behar du: 256, 512, 1024, 2048, 4096, 8192, 16 384. Balio baxuagoak erabiliz gero latentzia baxuagoa izango dugu, lagin-bloke txikiagoak bidaltzen baitira, maizago. Hala ere, erabiltzailearen ordenagailuaren prozesaketa-gaitasunaren arabera, audio-etenak eta jauziak gerta daitezke. Zenbait probaren ondoren, ikusi dugu neurri ona izan daitekeela 1024 edo 2048.

Audio-laginak *wav* fitxategi batean gorde daitezke, bai *Node.js* zerbitzarian, bai *AhoSR*n, hala etorkizuneko lanetarako baliagarria izan daitekeen materiala biltzeko asmoz.

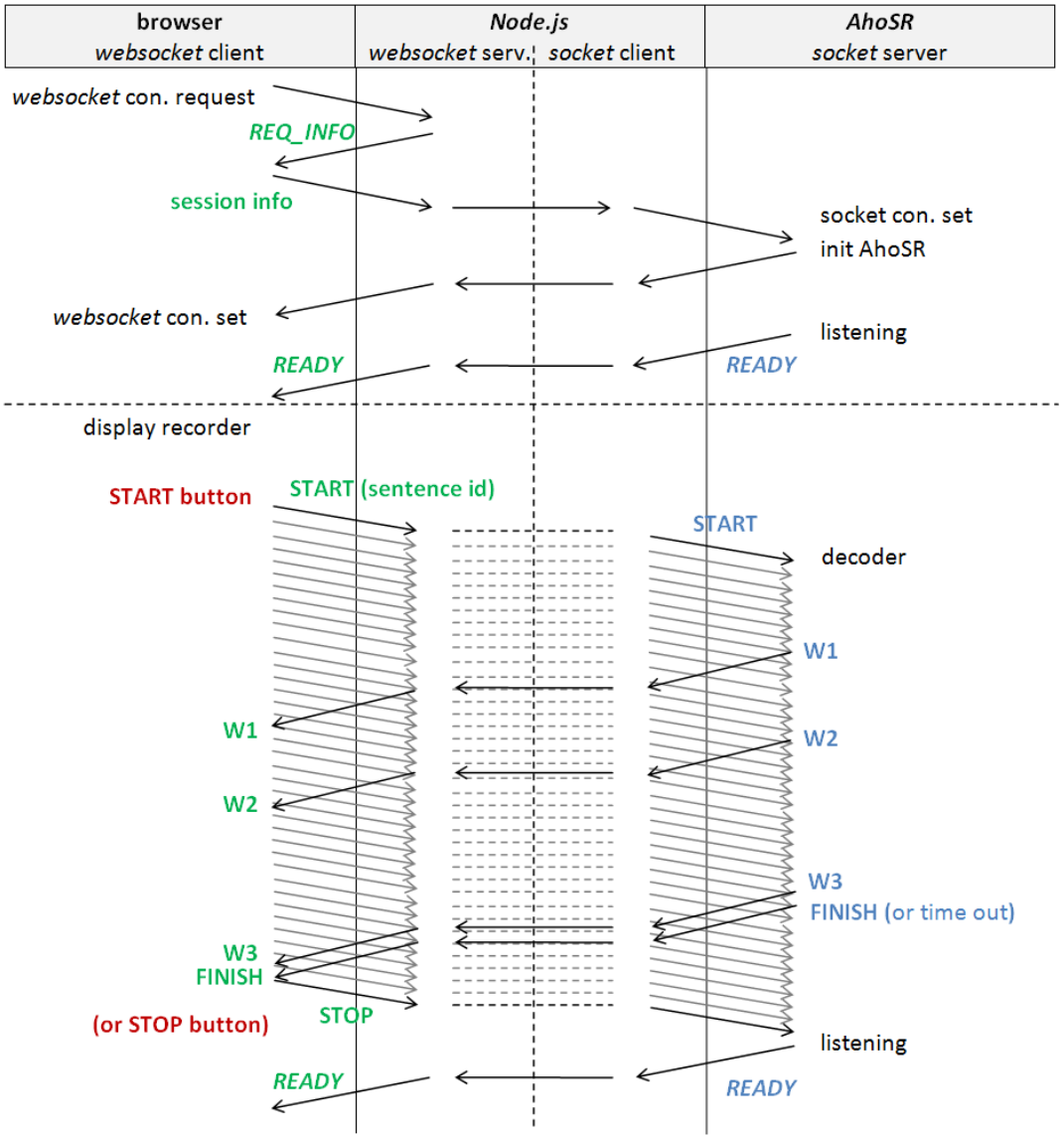


Figure 7.1: HHEE sistemako hiru atal osagaien arteko komunikazio-protokoloa denboran zehar (y ardatza): nabigatzailea (ezkerrean), *Node.js* zerbitzaria (erdian) eta *AhoSR* (eskuinean).

7.3 Konklusioak

Kapitulu honek azaltzen du zer egokitzapen behar diren gure sistema zerbitzari batean inplementatzeko, sarbide unibertsalaz. Bestalde, elkarrekin lanean jarri dira HTML5 berriaren espezifikazioetako bi API. *Audio API*ak mikrofonotik datuak grabatzeko

aukera ematen du, gero zerbitzarira bidaltzeko. OBELerako, audioa fitxategi batean paketatuta bidal daiteke, erabiltzaileak audioa egiaztatu ondoren; HHEErako, berriz, *websocket API*a baliatzen dugu, socket moduko konexio bat ezartzeko aukera ematen baitu nabigatzailearen eta zerbitzariaren artean. Mikrofono bidez jasotako audioa, hala, jaso ahala bidal daiteke, eta *feedback*a jaso.

CHAPTER 8

Online VADa

8.1 Sarrera

Ahots-aktibitatea detektatzea (VAD, *Voice Activity Detection*) gai garrantzitsua da ASR edo ASRan oinarritutako sistemetan. VADa erabiliz gero, audio-seinaleak hizketa-segmentu autonomotan zatitu daitezke, ondorengo moduloetara pasarazi aurretik. Hizketa-bilbeak soilik pasaraziz, ezagutzailearen konputazio-kostea txikiagotu egiten da, baita, horren ondorioz, deskodetze-prozesuaren erantzun-denbora ere [143]. Tesi honetan aurkeztutako sistemarentzat eta antzekoentzat, unibertsala izan behar du sarbideak; hortaz, VADak zarata-maila desberdinei aurre egiteko gai izan behar du, zehaztasun galerarik gabe —edo galera txikiekin—. Berez, gaur egungo sistemen erronkarik handiena da sarrerako hizketa-seinaleko hondoko zaratari aurre egitea [144].

Bi errore mota hartu behar dira kontuan: hizketa gisa pasatzen diren isiltasun- edo zarata-segmentuak (*isiltasunaren errore-tasa*) eta isiltasun gisa oker klasifikatzen diren hizketa-segmentuak eta, beraz, prozesaketa-sistemara heltzen ez direnak (*hizketaren errore-tasa*). Bi erroreei eutsi behar zaie baxu, jakina, baina bakoitzaren garrantzia VADa behar duen sistemaren diseinuaren beharrak baldintzatuko du.

VAD, eskuarki, prozesaketa akustikoko sistemen lehen modulua izan ohi da. Oso erabilia da mota guztietako sistema adituak garatzeko. [145]en, egileek ASR teknologia erabiltzen dute VAD batekin, motozikleta-ingurune batean elkarrizketa-sistema bat garatzeko. [146]en, larrialdietako ahots-komandoak prozesatzeko sistema integratu bat aurkezten da, non VADa erabiltzen den ASR baten aurretik. VAD eta ASR teknologiak jolas serioak dituen sistema baten hizketa-interfazearen muina ere badira, nahasmendu mentalentzako terapietarako integratuak [147]. Sistema horiek guztiak ASRan oinarrituak dira, eta, beraz, komeni da galdutako hizketa-bilbeen kopurua oso baxua izatea, erabiltzaileak erabilgarri izan ditzan audio-bilbe garrantzitsu guztiak. Bestalde, isiltasun-segmentuak hizketa gisa pasatuko balira, ezagutzaileak aukera du, hala ere, haiek hautemateko, zeren eskuarki isiltasunaren (edo ez-hizketaren) eredu bat izaten baitute. ASR interfazeetako VADaren helburu nagusia da isilune luzeak ezabatzea

eta audioa segmentu laburrago eta erabilgarriagotan zatitzea. Bide batez, konputazio-denbora murriztu ere egiten da, baita, hala, deskodetzearen erantzun-denbora.

ASRa ez da VAD modulu on bat behar duen teknologia bakarra. [148]en, hizlaria ezagutzeko arloko ikerketa-lerroetariko bat bezala identifikatzen dute VADa. Adibidez, atari-sistema adimendun batean erabiltzen da, non pertsonak beren ahotsaren bidez identifikatzen diren etxera sartzen utzi aurretik [149]. VADk ere modulu garrantzitsuak dira hizlariak segmentatzeko eta taldekatzeko sistemetan, hala nola [150]en aurkeztutako diarizazio-sisteman. Gainera, VADa funtsezko modulua da emozioak hautemateko sistemetan ere [151]. Hilaria eta emozioa ezagutzeko sistemetarako, komeni da oker klasifikatutako isiltasun- edo zarata-bilbe kopurua oso txikia izatea, zeren isiltasun- edo zarata-bilbeek ez dute emozioari edo hizlariaren identitateari buruzko informaziorik gordetzen. Isiltasunaren errore-tasa altua baldin bada, okerragotu egingo da sistemaren funtzionamendua. Bestela, ordea, zenbait hizketa-bilbe galtzen badira, sistemak behar bezala funtzionatzen jarrai dezake.

Gaur egungo VADak doitu egin daitezke modu batetik edo bestetik hurbilago jokatzeko. Hala ere, jokaera ideala litzateke ahalik eta gehien murriztea, bai isiltasunaren errore-tasa, bai hizketaren errore-tasa.

Sistema baten ahozko interfazeak audio-seinaleak gailu desberdinen bidez eta ingurune desberdinetan jasotzen dituenean, VADak grabazio-baldintza, kanal-ezaugarri eta zarata-maila desberdinei aurre egin behar die. Hori da, izatez, gaur egungo ASR sistemen erronkarik handiena [144]. Gaur egun, VAD sistemek egokitu egiten dituzte parametroak, hondoko zarata aldakorreko baldintzetara moldatzeko. Hala ere, teknika horrek baditu desabantailak: batetik, hasieratze-denbora bat behar da segmentu batean parametroak egokitzeko, eta horrek kaltegarria izan daitekeen atzerapena eragin dezake. Bestetik, parametroak ez badaude zuzen estimatuta, ezin da jakin sistemaren jokaera zein izango den [152]. VADa aurrez entrenatzea hasierako egokitzapena saihesteko modu bat da, baina entrenatutako sistemak gai izan behar du ikusi gabeko kanaletara edo hondoko zaratetara orokortzeko. VADaren *online*ko erabaki-hartzea erronka handia da oraindik ere.

Parametro akustikoen ikuspuntutik, oso parametro desberdinak ikertu izan dira: periodikotasun-neurria [153, 154], zero-gurutzatzeen tasa [155], tonua [156], epe laburreko energia (STE, *Short Term Energy*) [157] eta epe luzeko energia (LTE, *Long Term Energy*) [158, 159], espektro-analisia [160, 161], distantzia cepstrala [162], kodetze lineal prediktiboa (LPC, *Linear Predictive Coding*) [163] eta hainbat parametroren konbinazioak [164]. Zenbait ikerketa, oraintsuago, parametro multzoak erabiltzen saiatu dira ikasketa automatikoko teknikak erabiliz eredu estatistiko bat edo klasifikatzaile bat entrenatzeko, parametro akustiko diskriminatzaile berriak aztertu bainoago, zeina joera tradizionala baitzen.

Bai gaussiar nahasteen ereduak (GMM, *Gaussian Mixture Models*) eta Markoven ezkutuko ereduak (HMM, *Hidden Markov Models*) ere erabili izan dira VADaren testuinguruan. [165]en, hizketa- eta isiltasun-segmentuak bi HMMren bidez modelatzen dira.

Gramatika sinple bat erabiltzen da HMM batetik besterako trantsizioak modelatzeko, eta ahots-detekzioa, hala, ezagutze-sare batean zehar biderik onena aurkitzean datza. Erakusten dute HMMetan oinarritutako VAD sinple batek behar bezala funtzionatzen duela seinale garbiak kontsideratzen direnean. [166]en, HMM estrategia berari jarraitzen diote hondoko zaratari aurre egiteko, baina parametro akustikoak eta normalizazioa erabiltzen dira HMMen emaitzekin batera. [167]en, hainbat HMM zaratatsu entrenatzen dira isiltasun-segmentu zaratatsu desberdinak hautemateko. Tesi honetan, HMMek sortutako puntuazioen teknika darabilgu guk ere.

[168]en, urrutiko eremuko hizlariek gizakiaren eta makinaren arteko ahozko interakzioan sortzen duten interferentziaren arazoa lantzen da. Erabaki-zuhaitz (DT, *Decision Tree*) bat entrenatzen dute hizketa/isiltasun HMMen puntuazioak eta urrutiko eremuko hizlariari buruzko informazio gehigarria erabiliz. Bektore-euskarridun makina (SVM, *Support Vector Machine*) erabiltzen dute [169]en hizketa eta ez-hizketaren artean bereizteko, eta bertsio hobetuetan seinale/zarata ratioa (SNR, *Signal to Noise Ratio*) eransten dute [170, 171]. SVM/HMM arkitektura hibridoak ere proposatu dira VADrako [172]en, SVMaren propietate diskriminatzaile eta ez-linealei eusteko, bilbeen arteko korrelazioa HMM baten bidez modelatzen den bitartean. Emaitzek funtzionamendu hobea erakusten dute SVMan oinarritutako VAD sistemarentzat. Hala ere, hizketaren errore-tasek erlatiboki altu diraute. Guk proposatzen dugun VADak teknika horrek baino emaitza hobeak lortzen ditu, hizketaren errore-tasa hiru aldiz txikiagoz.

Oraintsuago, neurona-sareak (NN, *Neural Network*) agertu dira VAD metodoen literaturan. Adibidez, [173]ek neurona-sare errepikakorra (RNN, *Recurrent Neural Network*) erabiltzen du pertzepziozko predikzio linealeko (PLP, *Perceptual Linear Prediction*) parametroekin, seinale garbiak testatuz. Neurona-sare konboluzionalak (CNN, Convolutional Neural Networks) ere erabiltzen dira [174]en koefiziente mel-spectralekin, baina, ikusi gabeko kanaletarako, egokitzapena behar da datu ikuskatuekin. [175]en, log-mel iragazki-bankuko energietan oinarritutako parametro-bektoreek DT bat, SVM bat eta CNN klasifikatzaile bat elikatzen dute. Hala ere, VAD metodo horretan, doitu egin behar dira zenbait parametro, zarata-baldintza desberdinetara egokitzeko.

Online funtzionamenduari dagokionez, oraingo ikasketa sakoneko teknikek oso inferentzia-denbora luzeak behar dituzte; batez ere neurona-sareen arkitekturak ahalik eta konplexuenak izateko diseinatuak daudelako, denbora erreleko mugak kontuan hartu gabe [176]. Salbuespen bat da [177]en aurkeztutako sistema, non parametro akustiko desberdinak erabiltzen diren sinismen sakoneko neurona-sare bat (DBNN, *Deep Belief Neural Network*) entrenatzeko. Zarata mota desberdinak testatu ziren azterketa sakon batean lortutako emaitza esperimentalek adierazten dute erreferentziazko zenbait VADk baino emaitza hobeak dituela, baita denbora errealean ere. Hala ere, sistema horrek ia 300 parametro prozesatu behar ditu bilbe bakoitzean, eta horrek nabarmen handiagotzen du konplexutasuna. Aitzitik, gure teknikaz emaitza hobeak lortzen dira, eta sinpleagoa da.

Kapitulu honetan, normalizazio anitzeko puntuatzea (MNS, *Multi-Normalisation Scoring*) izena jarri diogun metodo batean oinarritutako VAD sinple baina izugarri

eraginkor bat aurkeztuko dugu. MNSren funtsa da isilune bateko audio-segmentuei dagokien Mel maiztasuneko koefiziente cepstral (MFCC, *Mel-Frequency Cepstral Coefficient*) normalizatuez entrenatutako HMM batek sortutako hainbat behaketa-egiantz klasifikatzea. Aurrez entrenatutako klasifikatzaile bat baliatzen du gure VAD teknikak; beraz, klasifikazio-ataza bat besterik ez da behar hizketa-bilbe berri bat iristen denean. Horrek esan nahi du emaitzak *online* lortzen direla, bilbez bilbe, eta ez dagoela parametro bat ere doitu beharrik; hortaz, ez da hasieratze-denborarik behar. Are gehiago, gaur egungo bi ITU-T VAD algoritmo estandarrekin alderatuz, gure VADak frogatu du askoz funtzionamendu hobea duela isiltasun-bilbeak etiketatzen, eta antzeko emaitzak lortzen dituela hizketa-bilbeak etiketatzen. Hori, konputazio-denbora luzatu gabe. VADa hainbat zarata motatarako testatu da, bai eta zenbait SNRtarako ere. Emaitzek agerian uzten dute guk proposatutako VAD teknika badela orokortzeko gai.

Kapitulua honela dago antolatuta: 8.2. atalean, isiltasun GMM batez lortutako behaketa-egiantzen zenbait alderdi aztertzen dira. Lehendabiziko esperimentu bat ere azaltzen da, zeinak erakusten baitu puntuazio horiek baliagarriak direla isiltasun- eta hizketa-segmentuen artean bereizteko. 8.3. atalean, artikulu honetan proposatzen den VAD sistemaren arkitektura orokorra deskribatzen da. 8.4. atalak MNS metodoa eta haren zergatia deskribatzen du. 8.5. atalean, erabilitako datu-baseen azalpen labur bat ageri da. VAD berriaren funtzionamendua ebaluatzeko, zenbait datu-base hautatu dira hainbat testuingururentzat baliagarria dela frogatzeko, test-material kopuru handia erabiliz. Zenbait esperimenturen emaitzak (bai baldintza garbietan, bai zaratatsuetan) 8.6. atalean ageri dira. 8.7. atalean, balidazio-esperimentu bat deskribatzen da, emaitzak bi VAD sistema estandarrenekin alderatuz; azkenik, zenbait konklusio azaltzen dira 8.8. atalean.

8.2 Behaketa-egiantza

Hizketa-ezagutzan, ezagutze-unitate berari dagokien audio-segmentuak (hitza, fonema, trifonema eta abar) batu eta prozesatu egiten dira, haietatik parametro akustikoak erauzteko (eskuarki, MFCCak) eta unitate bakoitzerako eredu akustiko desberdin bat entrenatzeko. HMMa oso eredu akustiko ezaguna da; izan ere, behaketa-bektore berri baten egiantza modelatu ez ezik, behaketen sekuentzialtasuna ere modelatzen du.

Behaketa-egiantzak GMMek sortzen dituzte, eta GMM bakoitza HMM egoera bati dagokio. o_t behaketa-bektore batentzat, GMM baten b_j behaketa-egiantza j_{th} egoeran (8.1) ekuazioan adierazi bezala kalkulatzen da.

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} N\left(o_t; \mu_{jm}, \Sigma_{jm}\right) \tag{8.1}$$

non M: nahaste-osagaien kopurua; c_{jm}: m^{th} osagaiaren pisua; $N(\cdot; \mu, \Sigma)$: aldagai anitzeko gaussiarra, μ batezbesteko-bektorea eta Σ kobariantza-matrizea dituena.

Lan honetan, *Basque Speecon-like* datu-basea erabiliz (bereziki *close-talk* kanala) entrenatutako isiltasun HMMtik lortu dira behaketa-egiantzak [99],

8.2.1 Isiltasunaren eredu akustikoa

Isiltasun bilbeetarako hautatutako HMM topologiak hiru egoera ditu eta ezkerretik eskuinerakoa da, baina eskuineko egoera ezkerrekoarekin loturik du. Parametro akustiko gisa 13 MFCC eta lehen mailako 13 eta bigarren mailako 13 deribatu erabili dira HMMa entrenatzeko, eta 32 nahasteko GMMak. Bilbearen luzera 25 ms da, 10 ms oro.

MFCCei CMVN aplikatu zitzaien, eta batezbesteko eta bariantza orokorrak erauzi ziren grabazio-saio bakoitzetik. N bektore cepstralentzat $y = \{y_1,y_2,...,y_N\}$, haien μ_N batezbesteko-bektorea eta σ_N^2 bariantza-bektorea (8.2) ekuazioan eta (8.3) ekuazioan ageri den bezala kalkulatzen dira, hurrenez hurren.

$$\mu_N(i) = \frac{1}{N}\sum_{n=1}^{N} y_n(i) \tag{8.2}$$

$$\sigma_N^2(i) = \frac{1}{N}\sum_{n=1}^{N} (y_n(i) - \mu_N(i))^2 \tag{8.3}$$

non i: bektorearen i^{th} osagaia den.

Gero, parametro cepstralak normalizatu egiten dira, kalkulatutako batezbesteko- eta bariantza-bektoreak erabiliz, (8.4) ekuazioan ageri den bezala. Hala, parametro normalizatu bakoitzak zero balioko batezbestekoa du, eta bat balioko bariantza.

$$\hat{y}_n(i) = \frac{y_n(i) - \mu_N(i)}{\sigma_N(i)} \tag{8.4}$$

8.2.2 CMVNaren eragina

CMVNaren erabilerak eragin nabarmena du behaketa-egiantzek osatzen duten kurbetan. Lagin-seinale bat testatzean isiltasun HMMaren egoera bakoitzeko behaketa-egiantzak bilbez bilbe kalkulatuta, oso kurba desberdinak lortzen dira, CMVN aplikatzen bada edo ez. 8.1. irudiak desberdintasun hori ilustratzen du. Erdiko eta beheko diagrametan, ageri da s_0, s_1 eta s_2 HMM egoeretan sortutako behaketa-egiantzen log-en kurbak, normalizaziorik gabe eta normalizazioarekin, hurrenez hurren, lau hitzez osatutako esaldi batean. Normalizazioa, kasu horretan, fitxategitik bertatik erauzitako batezbestekoak eta bariantzak erabiliz kalkulatu da.

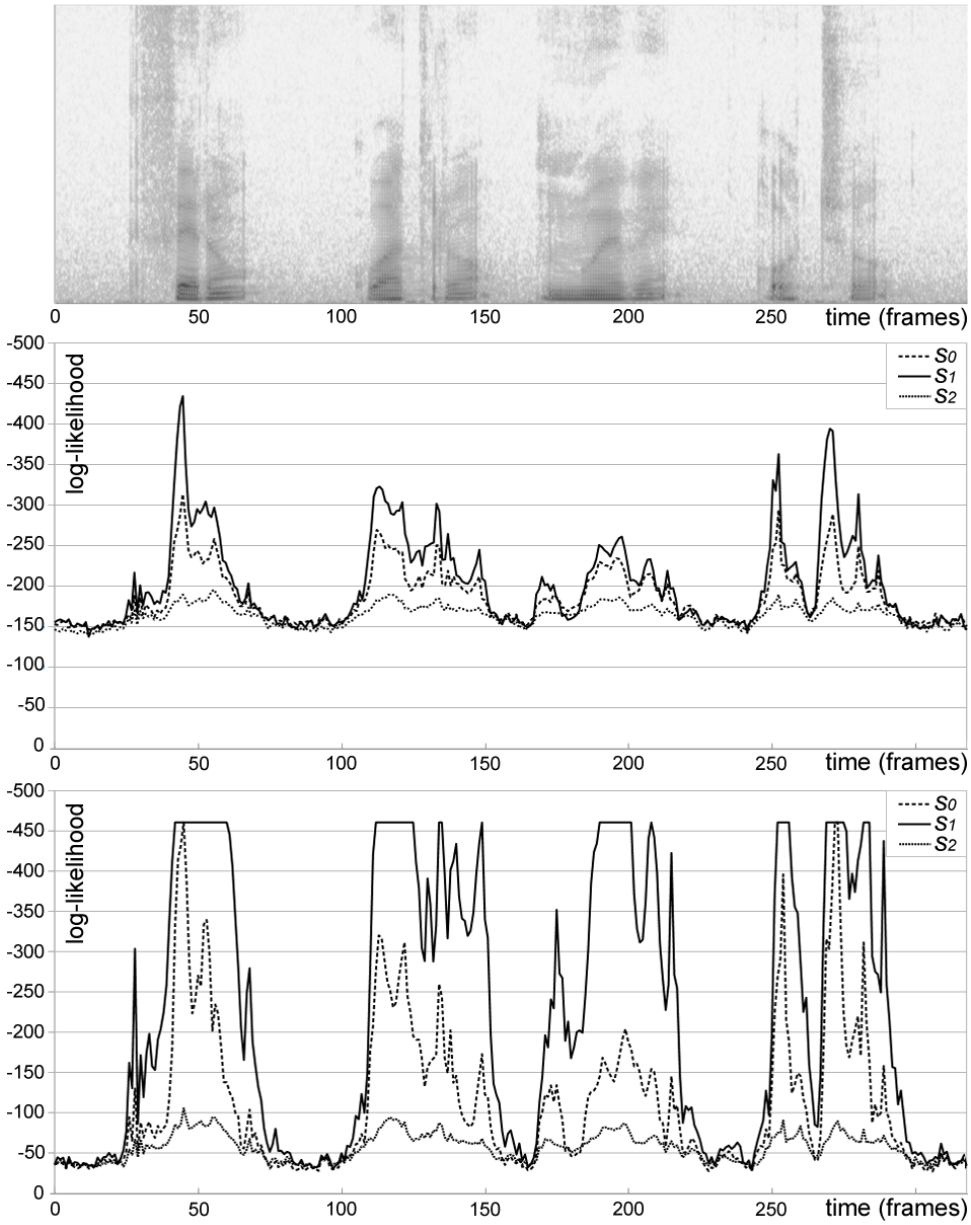


Figure 8.1: Espektrograma (goian) eta *Isiltasun* HMMaren ezkerreko egoerak (s_0), erdiko egoerak (s_1) eta eskuineko egoerak (s_2) sortutako behaketa-egiantzen log-a denboran (bilbeetan) zehar, CMVNrik gabe (erdian) eta CMVNarekin (behean).

Beheko diagramako kurbek (CMVNarekin), erdiko diagramakoekin kalkulatuta (CMVNrik gabe), pikoagoak dirudite. Ezaugarri hori hizketa eta ez-hizketa hobeki bereizteko erabil daiteke.

8.2.3 Isiltasun HMMaren erdiko egoera

Hiru egoerako HMM bateko erdiko egoera, berez, ereduaren egoerarik egonkorrena da; izan ere, muturretako egoerek ereduen arteko trantsizioak kudeatu behar dituzte. Zentzuzkoa dirudi pentsatzeak gauza bera gertatuko dela isiltasun HMMarekin, non muturretako egoerek isiltasunaren eta hizketaren arteko trantsizioak modelatu behar baitituzte.

Adibide adierazgarri bat 8.2. irudian dago. HMM egoera (s_0, s_1 eta s_2) bakoitzeko GMMak sortutako log-egiantzak ageri dira, hiru hitzeko esaldi batean zehar (erreparatu ezpain-soinuari, bigarren hitzaren aurretik). Erdiko egoerako (s_1) GMMak sortutako behaketa-egiantzen kurbak askoz diskriminatzaileagoa dirudi muturretako kurbak baino, zeinak irregularragoak baitira. Analisia egiteko, "*M25 — R+M25*" HMM multzoko isiltasun HMMa erabiliz egin da (ikus 5. kapitulua), *offline*ko CMVNaz (*close-talk* azpimultzoa erabiliz entrenatua, parametro akustiko gisa 13 MFCC eta lehen mailako 13 eta bigarren mailako 13 deribaturekin eta 32 nahasteko GMMekin).

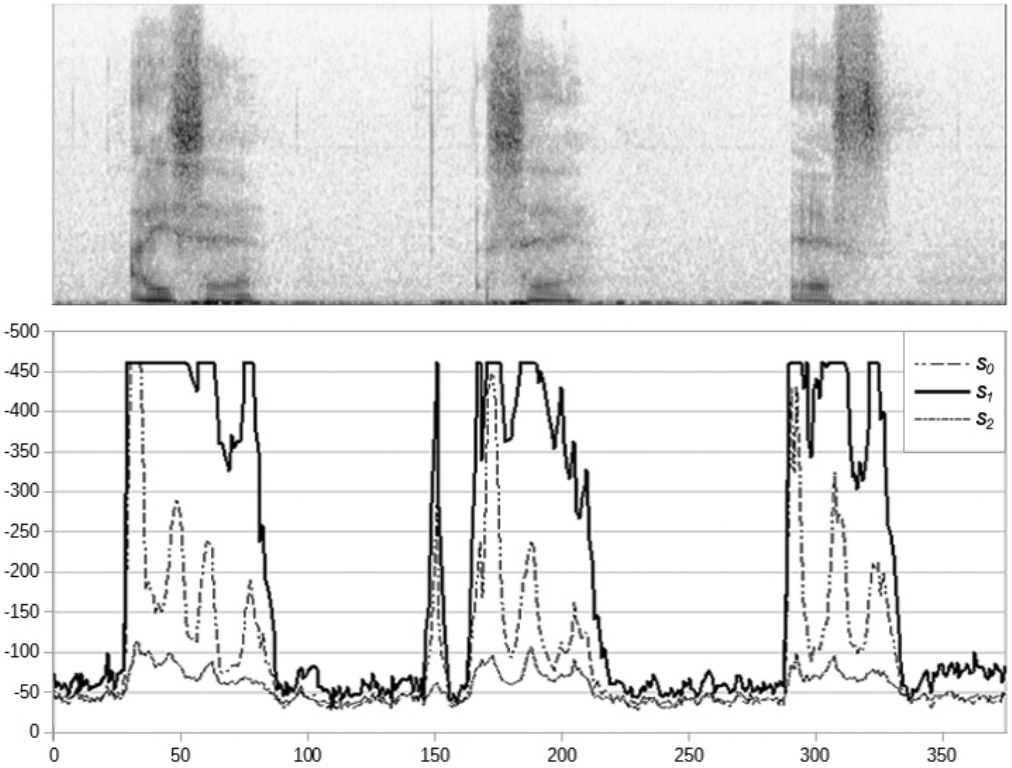


Figure 8.2: Hiru hitzez osatutako esaldi baten espektrograma (goian) eta MFCC normalizatuz entrenatutako isiltasun HMMko ezkerreko egoeran (s_0), erdiko egoeran (s_1) eta eskuineko egoeran (s_2) sortutako behaketa-egiantzen logak denboran (bilbeetan) zehar (behean).

Hasieran, egiaztatu genuen isiltasun HMMaren erdiko egoeraren GMMak sortutako egiantzek oso jokaera desberdina dutela, HMMaren entrenatzeko moduaren arabera. HMM desberdinak testatu ziren, eta kontua da emaitza nabarmen narriatzen dela entrenamendu-datuek akats gehiago eduki ahala. Jokaera diskriminatzailerik onena duen HMM multzoa —eta, beraz, emaitzarik onenak izatea espera dena— entrenamenduko hasierako egoeretan $M25$ azpimultzoa (eskuz etiketatua) erabiliz sortutakoa da. Hortaz, bistakoa da zer garrantzitsua den gure datu-basearentzat fase egokiak erabiliz entrenatzea. 8.3. irudian, HMM desberdinen erdiko egoerek sortutako log-egiantzak ageri dira; diskriminatzaileena kurba ilunena da ($M25$ — $M25{+}R$), irregularrena, aldiz, kurba argiena ($M25{+}R$).

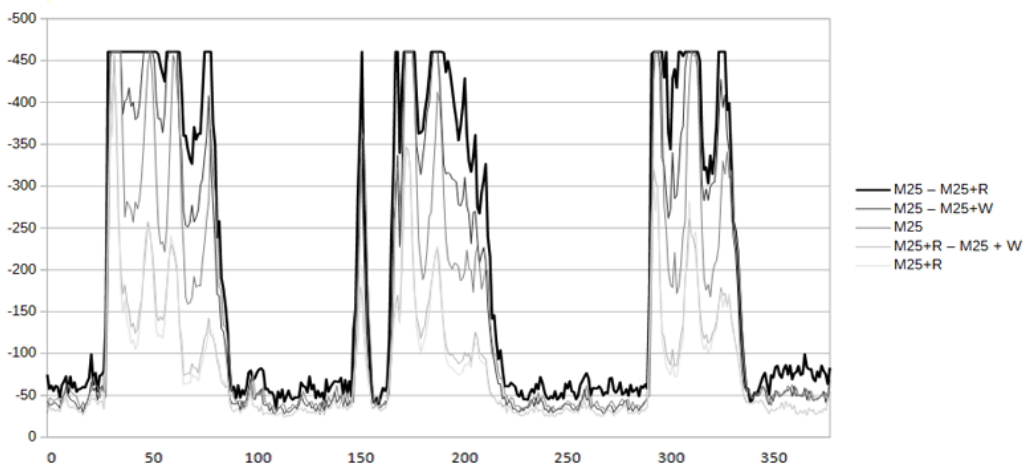


Figure 8.3: Cepstrum normalizatuak erabiliz hainbat modutan entrenatutako isiltasun HMMetako erdiko egoeretan (s_1) lortutako behaketa-egiantzen logak, denboran (bilbeetan) zehar.

Ahots-aktibitatearen detektatzaile batean azpimarratu beharreko beste interes-puntu bat zaratarekiko sendotasuna da. 8.4. irudian, lau zarata-baldintza desberdinetan grabatutako esaldi beraren egiantzak ageri dira (*Spanish SpeeCon*etik atereak dira, ikus 8.2.4. atala). Esaldia lau distantziatara grabatu zen, lau mikrofono desberdin erabiliz. Grabazio bakoitzean, kanal desberdin bat ageri da (C_0, C_1, C_2 eta C_3), SNR balio desberdinekoa: garbiena, 20 dB ingurukoa (C_0), eta zaratatsuena, 0 dB ingurukoa (C_3). Esaldiak zenbaki-segida hau du espainieraz: "*cero*, *cuatro*, *nueve*, *ocho*" (zero, lau, bederatzi, zortzi).

Probabilitate-kurbek erakusten dute ezen, espero bezala, okerrera egiten dutela seinale zaratatsuagoak prozesatzean, baina, orduan ere, aski diskriminatzaile jarraitzen dute kurbek. 0 dB-tan, efekturik kaltegarriena hasierako eta amaierako fonemetan gertatzen da, non, fonemaren arabera, probabilitateak isiltasun zaratatsuaren probabilitateen oso antzekoak izan baitaitezke. Hori, batez ere, hasierako fonema fonema zaratatsua denean

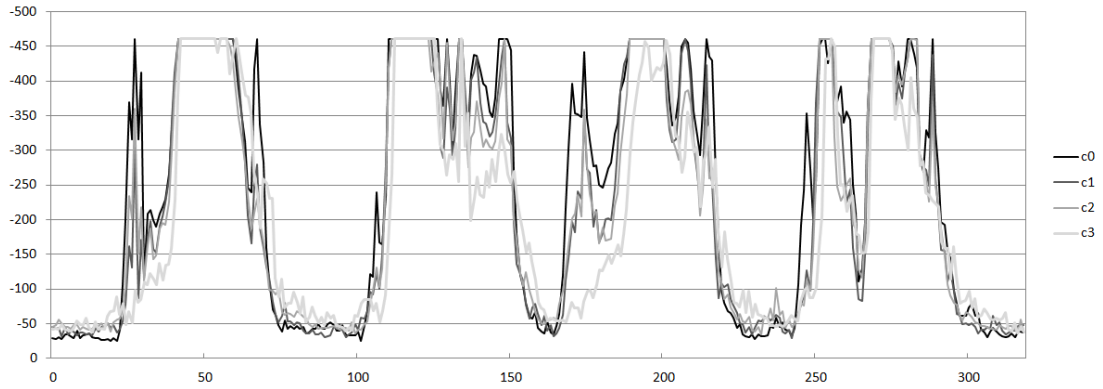


Figure 8.4: Isiltasun HMMaren erdiko egoeran (s_1) lortutako behaketa-egiantzen loga, denboran (bilbeetan) zehar, SNR desberdineko audio-seinaleak prozesatzean: C_0tik (20 dB) C_3ra (0 dB).

gertatzen da. Hala ere, VADak jokaera ona du gainerako baldintza ez hain zaratatsuetan, eta probabilitate-profilak oso antzekoak dira 20 dB-ko SNRaz lortutakoarekin alderatuta.

8.2.4 Hasierako VAD esperimentua

Isiltasun HMMaren erdiko egoerak sortutako behaketa-probabilitateen kurben egonkortasuna balioesteko, VAD zehaztasun-esperimentu bat egin da. Fitxategi bakoitzaren parametroak normalizatu egin dira, fitxategitik bertatik kalkulatutako batezbestekoak eta bariantzak erabiliz; alegia, kontsideratzen dugu *offline*ko VAD esperimentua dela, fitxategi bakoitza aurrez prozesatu baita bilbe bakoitza banan-banan klasifikatzen hasi baino lehen. Cepstrumak *online* kalkulatzeko, hasierarako balioak estimatu behar dira (ikus **??**. atala), eta estimazio horrek izugarrizko eragina izan dezake VADaren funtzionamenduan. Hala ere, hurrengo atalean azaltzen den MNS metodoaz konponbide eraginkorra ematen zaio arazo horri.

Esperimenturako, aurreko ataletan testatutako isiltasun HMMrik diskriminatzaileena hautatu da: "*M25 — R+M25*" isiltasun HMMa. Testatutako fitxategiei dagokienez, *Spanish SpeeCon* datu-basearen azpimultzo bat hautatu da [178], aurreko ikerketa-lan batean VAD algoritmo desberdinak ebaluatzeko erreferentzia gisa erabili izan dugu eta [179]. 1000 esaldi baino gehiago ditu azpimultzo horrek, hainbat ingurunetan grabatuak (bulegoa, aisialdi-lekuak, automobila eta leku publikoak). Esaldi bakoitza lau mikrofono desberdinen bidez grabatu zen: hurbileko entzungailu baten mikrofonoa (C_0 kanala), paparreko mikrofonoa (C_1 kanala), distantzia ertaineko mikrofono kardioidea (0.5-1 m, C_2 kanala), eta norabide orotako urruneko mikrofonoa (C_3 kanala). Hala, horietariko kanal bakoitzak ingurune desberdinetan grabatutako seinaleak ditu. Eszenatokirik garbiena C_0 da; zaratatsuena, berriz, C_3. Datu-baseko seinalek 16 kHz-eko laginketa-maiztasunez grabatu ziren, lagin bakoitzeko 16 bit erabiliz. Erreferentziazko hizketa eta

isiltasun-etiketak aipatutako hasierako lanean erabilitakoak dira.

VAD zehaztasun-esperimentuaren funtsa da sistemak hizketa- eta isiltasun-segmentuak bereizteko duen gaitasuna ebaluatzea lau SNR desberdinetan, isiltasunaren errore-tasa (ER_0) eta hizketaren errore-tasa (ER_1). Bi tasa horiek honela kalkulatzen dira: oker klasifikatutako isiltasun- edo hizketa-bilbeak ($N_{0,1}$ eta $N_{1,0}$, hurrenez hurren) zati datubase osoko benetako isiltasun- eta hizketa-bilbeak (N_0^{ref} eta N_1^{ref}, hurrenez hurren), (8.5) ekuazioan ageri denez. Gainera, errore-tasa osoa (TER, *total error rate*) ere kalkulatu da, honela: oker klasifikatutako bilbeen kopurua zati bilbe kopuru osoa ((8.6) ekuazioa).

$$ER_0 = \frac{N_{0,1}}{N_0^{ref}} \times 100; ER_1 = \frac{N_{1,0}}{N_1^{ref}} \times 100 \tag{8.5}$$

$$TER = \frac{N_{0,1} + N_{1,0}}{N} \times 100 \tag{8.6}$$

Esperimentua atalase desberdinetarako eta hizketa- eta isiltasun-segmentuen iraupen desberdinetarako egin da. Emaitzarik onenak bai hizketa-segmentuetarako, bai isiltasun-segmentuetarako 15 bilbeko gutxieneko iraupena jarrita lortu dira, eta hala lortutako emaitzak 8.5. irudian ageri dira. TER emaitzarik onenak $Th = -150$ atalaseaz lortu dira hainbat SNR mailatan. Gainera, nahiko segmentu laua dago kurbetan, -150 eta -200 artean, eta horrek baieztatzen du badela funtzionamendu-tarte egonkor bat. ER_0 eta ER_1 kurben errore-tasa berdineko (EER, *Equal error rate*) puntuak bi balio horien artean kokaturik daude.

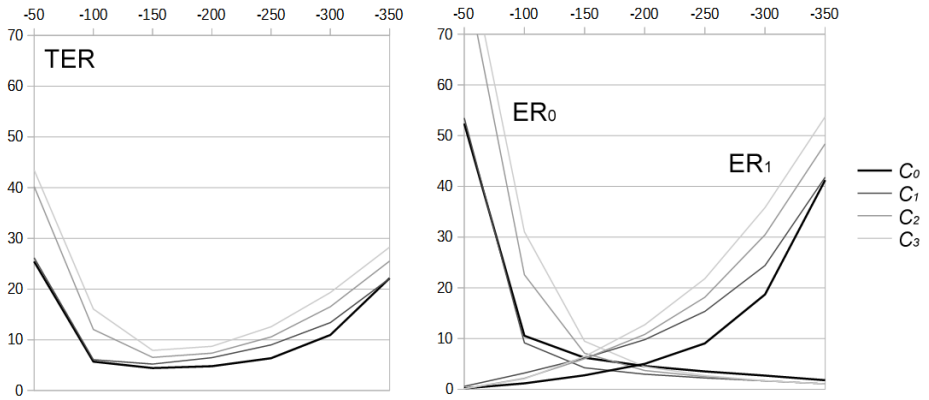


Figure 8.5: *Offline*ko VAD zehaztasun-esperimentua: TER (ezkerreko irudia) eta ER_0 eta ER_1 (eskuineko irudia) lau SNR mailetan, hainbat atalase-baliotarako.

8.1. taulan, $Th = -150$ atalaserako lortutako emaitzak ageri dira, lau SNR mailetan. Emaitzek azaltzen dute ezen, espero bezala, errore-tasek gora egiten dutela SNR jaitsi ahala. Hala ere, isiltasunaren errore-tasarik onena C_1ean lortzen da.

Table 8.1: *Offline* esperimentuko TER, ER_0 eta ER_1, $Th = -150$ atalaserako eta lau kanal desberdinetarako.

	TER	ER_0	ER_1
C_0	4.42	6.21	2.74
C_1	5.21	4.22	6.13
C_2	6.53	7.10	6.00
C_3	7.90	9.46	6.45

Hizketa-prozesaketarako, garrantzitsua da ER_1 ahal bezainbat murriztea. Horretarako, hizketa-segmentuei 10 bilbeko marjina gehigarria ezarri zaie, zeren segmentuen mugetan agertzen baitira hizketa-sailkapenaren errore asko, batez ere fonema zaratatsuez hasten diren hitzetan. Horrekin, nabarmen jaisten da ER_1, TERean eragin handirik izan gabe. Kontuan izatekoa da mugak ez direla inoiz zehatzak; hori dela eta, zenbait ebaluaziotan, segmentuen inguruko ± 5 ms-ko tarteak baztertu egiten dira ebaluazio-prozesutik [180]. 8.2. taulan, beste esperimentu horren emaitzak ageri dira, non, datu-multzo bera [179] erabiliz, proposatutako teknikarekin lortutako emaitzak eta lau VAD algoritmo estandar ezagunenak ere alderatzen baitira: G.729 sistema [155], FD (frame-dropping mechanism) eta NR (noise reduction system) algoritmoak AFE-DSRn (advanced frontend for distributed speech recognition) inplementatuak [181] eta LTSE (long-term spectral divergence) algoritmoa [182].

8.2. taulak erakusten du errore-tasa osorik onena, hizketa eta isiltasuna batera hartuta, guk proposatutako teknikarekin lortzen direla. Hizketari dagokionez, AFE-FD eta LTSE sistemek emaitza hobeak lortzen dituzte; halere, badute desabantaila bat: isiltasun-sailkapenean oso errore-tasa handiak dituzte kanal guztietan (emaitzarik baxuena % 38.10 da). Horrek esan nahi du isiltasun-bilbe asko bidaliko liratekeela ezagutzailera. Gure hasierako sistemak % 12.42 eta 17.59 bitarteko errore-tasak ematen ditu, isiltasun-bilbeen sailkapenean.

8.2.5 Konklusioak

Laburbilduz, emaitzek erakusten dute ezen isiltasun HMMaren erdiko egoerako GMMak duen bereizte-gaitasuna, oso sinplea izanik ere, oso ona dela. Lortutako errore-tasa osoak, gutxi gorabehera, testatutako bigarren sistema onenaren erdia dira. Hala ere, tasarik garrantzitsuena hizketa-bilbeen errore-tasa da (ER_1), zeren ezagutzailerantz pasatzen uzten diren isiltasun-bilbeak oraindik ere isiltasun gisa sailkatu daitezke deskodetzailean. Gure sistemak oso errore-tasa baxuak ateratzen ditu kanal guztietan, altuena % 2.89 izanik, mikrofono urrunenaren kanalerako (C_3).

Table 8.2: Hainbat VAD algoritmoz lortutako emaitzen konparaketa, lau SNR mailatan

(c) Errore-tasa osoa (TER)

	G.729	AFE-FD	AFE-NR	LTSE	Prop.
C_0	28.98	30.49	28.11	18.68	7.99
C_1	38.74	26.24	27.73	16.22	7.20
C_2	38.16	25.09	20.69	19.02	8.71
C_3	42.94	24.61	27.05	17.54	9.99

(b) Isiltasunaren errore-tasa (ER_0)

	G.729	AFE-FD	AFE-NR	LTSE	Prop.
C_0	56.06	63.88	63.88	58.23	15.68
C_1	70.23	54.75	54.75	55.96	12.42
C_2	59.54	59.54	52.10	38.10	15.39
C_3	70.49	70.49	50.10	47.65	17.59

(a) Hizketaren errore-tasa (ER_1)

	G.729	AFE-FD	AFE-NR	LTSE	Prop.
C_0	3.63	0.03	0.62	0.05	0.79
C_1	9.28	0.23	1.98	0.49	2.30
C_2	18.19	0.48	4.83	0.53	2.47
C_3	17.22	1.41	8.30	1.34	2.89

8.3 MNSan oinarritutako VAD sistemaren arkitektura orokorra

Tesi honetan proposatutako *online* VAD teknikak hiru bloke nagusi ditu, 8.6. irudian ageri denez. Sistemaren sarrera MFCCen bektore bat da, uneko seinale-bilbetik lortzen dena; irteera, berriz, VAD etiketa bat da: *hizketa* edo *isiltasuna*.

Hiru bloke nagusiak hauek dira:

- Normalizazio cepstralaren modulua: sarrerako seinale-bilbearen parametro akustikoak (MFCCak) hainbat normalizazio-faktore erabiliz normalizatzen dira.
- MNSan oinarritutako MLP modulua (VAD sailkatzailea): guk proposatutako MNS metodoaz lortutako bektore bat sailkatzen du modulu horrek, geruza anitzeko pertzeptroia (MLP, *Multi-Layer Perceptron*) erabiliz.
- Modulu erabaki-hartzailea: berehalako erabakiak hartzeko egoera finituko automata bat du; hala, zarata laburrak maneiatzen dira, eta emaitzak hobetu.

Hemen aurkeztutako metodoa 2. blokean dago inplementatuta, eta kapituluan zehar

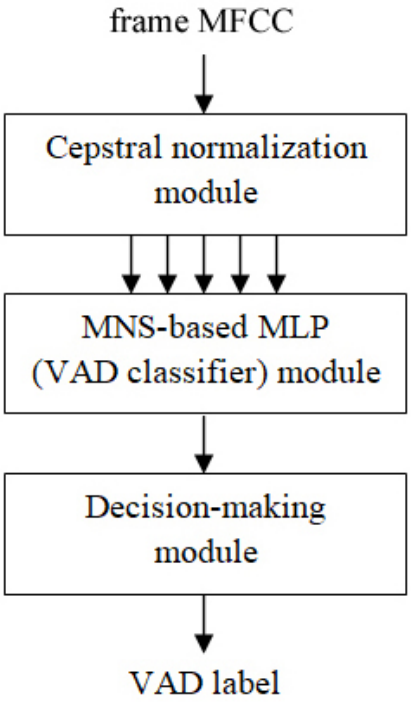


Figure 8.6: Proposatutako *online* VAD teknikaren arkitektura orokorra.

dago deskribaturik. 1. eta 3. blokeak hurrengo azpiataletan deskribatzen dira xeheago.

8.3.1 Normalizazio cepstrala

Normalizazio cepstrala ezinbestekoa da lan honetan proposatutako VADa garatzeko. Berez, lehenagoko lanetan frogatu denez [183], MFCC normalizatuekin entrenatutako isiltasun GMMaz lortutako behaketa-egiantzek nahiko patroi bereiziak jarraitzen dituzte hizketarako eta isiltasunerako. Lan honetan proposatutako VADak ezaugarri hori baliatzen du.

Oro har, parametro-normalizazioa nahitaezkoa da eredu akustiko sendoak sortzeko eta ingurune desberdinetan jasotako audio-seinaleak maneiatzeko. [184]ko kenketa espektralaren ikuspuntua aski ezaguna da ASR-ren eremuan, sarrerako seinaleen desberdintasunak konpentsatzeko (kanalak, hondo-zarata eta abar). Hala ere, praktikarik erabiliena da erauzitako parametroetan CMVN aplikatzea, kenketa espektralak baino emaitza hobeak baititu [185].

[133]n azaltzen denez, N bilbetan zeharreko MFCC batezbestekoak uneko mikrofonoaren eta gelaren akustikaren ezaugarri espektralak barne hartzen ditu. Limitean, $N \longrightarrow \infty$ doanean, espero daiteke ingurune berean jasotako grabazio desberdinen batezbestekoak berbera izatea. Hala, batezbesteko normalizazio cepstrala (CMN, *Cep-*

stral Mean Normalisation) lagungarria da kanalaren transferentzia-funtzio geldikor eta lineala ezabatzeko; bariantzaren normalizazioa (CVN), berriz, zarata gehigarriaren ondorioz MFCCek duten bariantza-murrizketa konpentsatzeko.

CMVN metodo klasikoan [186, 187] parametro cepstral (MFCC) bakoitzeko batezbesteko eta bariantza-bektoreak estimatzen dira. Parametro-bektoreak, orduan, lerratu eta eskalatu egiten dira estimatutako batezbestekoak eta bariantzak erabiliz, eta, ondorioz, parametro normalizatu bakoitzak zero balioko batezbestekoa eta bat balioko bariantza du. Batezbesteko eta bariantza baliozkoak kalkulatzeko modu eraginkor bat da fitxategi osoa erabiltzea haiek kalkulatzeko (*offline* funtzionamendua). Fitxategian oinarritutako normalizazio horrek, baina, nahi ez diren atzerapenak eragin ditzake, zeren esakuntza ezin da prozesatzen hasi azken bilbea iristen den arte. Sistema sinkrono edo *online*koetan, 150-200 ms-ko gutxieneko luzera duten leihoak erabili ohi dira batezbesteko eta bariantza estimatuen kalitatearen eta eragindako atzerapenaren arteko erdibide gisa. Hasierako balioa estimatuta, normalizazio errekurtsibo motaren bat aplikatu ohi da.

Batezbestekoen eta bariantzen hasierako balioak lehen N bilbeak erabiliz estimatu daitezke (eta, ondoren, errekurtsiboki egokitu). Hasierako balio horien estimazioa oso bestelakoa izango da N bilbe horiek hizketa baldin badute edo ez. Bilbeetan hizketa baldin badago, kalkulatutako bariantzen balioak oso txikiak izango dira, eta, horren ondorioz, izugarri anplifikatuko da seinale normalizatuaren anplitudea; eta alderantziz. Beraz, berebiziko garrantzia du batezbestekoen eta bariantzen hasierako balioak ondo estimatzeak. Kontu hori konpontzeko, tesi honetan aurkeztutako metodo bat erabil daiteke, parametro cepstralei hainbat normalizazio-faktore aplikatzean datzana. Erabakiak bilbez bilbe hartzeko aukera ematen du horrek, leihorik erabili beharrik gabe.

8.3.2 Modulu erabaki-hartzailea

Hizketa/isiltasuna erabakia bilbez bilbe hartzen denez, hizketa gisa etiketatutako segmentu labur-laburrak ager daitezke VADaren irteeran. Segmentu labur horiek, eskuarki, zarata txikien ondorioz sortzen dira, eta eragina izan dezakete ondorengo prozesaketa-sistemaren funtzionamenduan. *Offline* inplementazioan, post-prozesaketa izan ohi da; *online* inplementazioan, ostera, berehalakoan hartu behar dira erabakiak. Gure *online* inplementazioan, egoera-diagrama klasiko bat inplementatu da (ikus 8.7. irudia). Bi parametro kontsideratu dira: hizketaren gutxieneko iraupena (T_{min_speech}) eta isiltasunaren gutxieneko iraupena (T_{min_sil}), zeinak segmentu batek hizketa eta isiltasun gisa jotzeko hurrenez hurren eduki behar dituen gutxieneko kopuruak ezartzen baitituzte. Irudian ikus daitekeenez, VADak bilbe jakin batean bere egoera isiltasunetik hizketara aldatzen badu (edo alderantziz), ondorengo T_{min_speech} bilbeak (or T_{min_sil}) ere analizatzen dira. Bilbe horien azterketaren emaitza bat baldin badatoz uneko bilbearenarekin, aldatu egiten da egoera; bestela, jotzen da zaratatxo bat izan dela eta VADk ez du bere egoera aldatzen. Bistan denez, metodo honek atzerapen txiki bat eransten du egoera-aldaketa bat aurkitzen den bakoitzean, baina ez da atzerapenik gehitzen egoera berari eusten zaion bitartean. Horrek segmentuen barnean bere onera

erabat bueltatzeko aukera ematen dio sistemari.

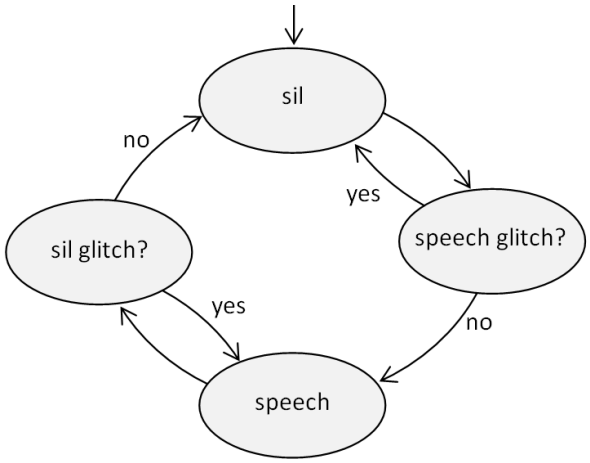


Figure 8.7: VADaren *online* inplementazioaren egoera-diagrama.

For the experiments carried out in this thesis, a minimum segment duration of 15 frames was empirically chosen for both T_{min_speech} and T_{min_sil}.

Tesi honetan egindako esperimentuetarako, 15 bilbeko gutxieneko segmentu-iraupena hautatu da enpirikoki, bai T_{min_speech}-erako, bai T_{min_sil}-erako.

8.4 MNS metodoa

MNS metodoaren funtsa da hainbat behaketa-egiantz sortzea MFCCak parametro desberdinekin normalizatuz, alegia, baldintza akustiko desberdinetan grabatutako hainbat datu-basetako batezbestekoak eta bariantzak erabiliz. Hala lortutako behaketa-egiantzen bektoreek hizketa- eta isiltasun-bilbeen jokaera karakteriza dezakete, baldintza desberdinetan. 8.8. irudian, adibide argigarri gisa, hurbiletik jasotako seinale baten puntuazioak (goian, C_0) eta urrutitik jasotako beste batenak (behean, C_3) ageri dira, normalizatuta, aldi berean lau distantziatan grabatutako lau datu-multzo erabiliz kalkulatutako batezbestekoekin eta bariantzekin: hurbilekoa (C_0), mahai-gainekoa (C_1), distantzia ertainekoa (C_2) and urrutikoa (C_3).

Jotzen badugu isiltasun-segmentuetako puntuazioek, modu desberdinetan normalizatuta, patroi bati eusten diotela (ikus 8.8. irudiko s_i puntuazio-bektorea), litekeena da hizketa-puntuazioek ere patroi bati eustea. Hala bada, sailkatzaile on bat besterik ez da behar patroi horiei antzeman eta bektorea hizketa-bilbe bati edo isiltasun-bilbe bati esleitzeko.

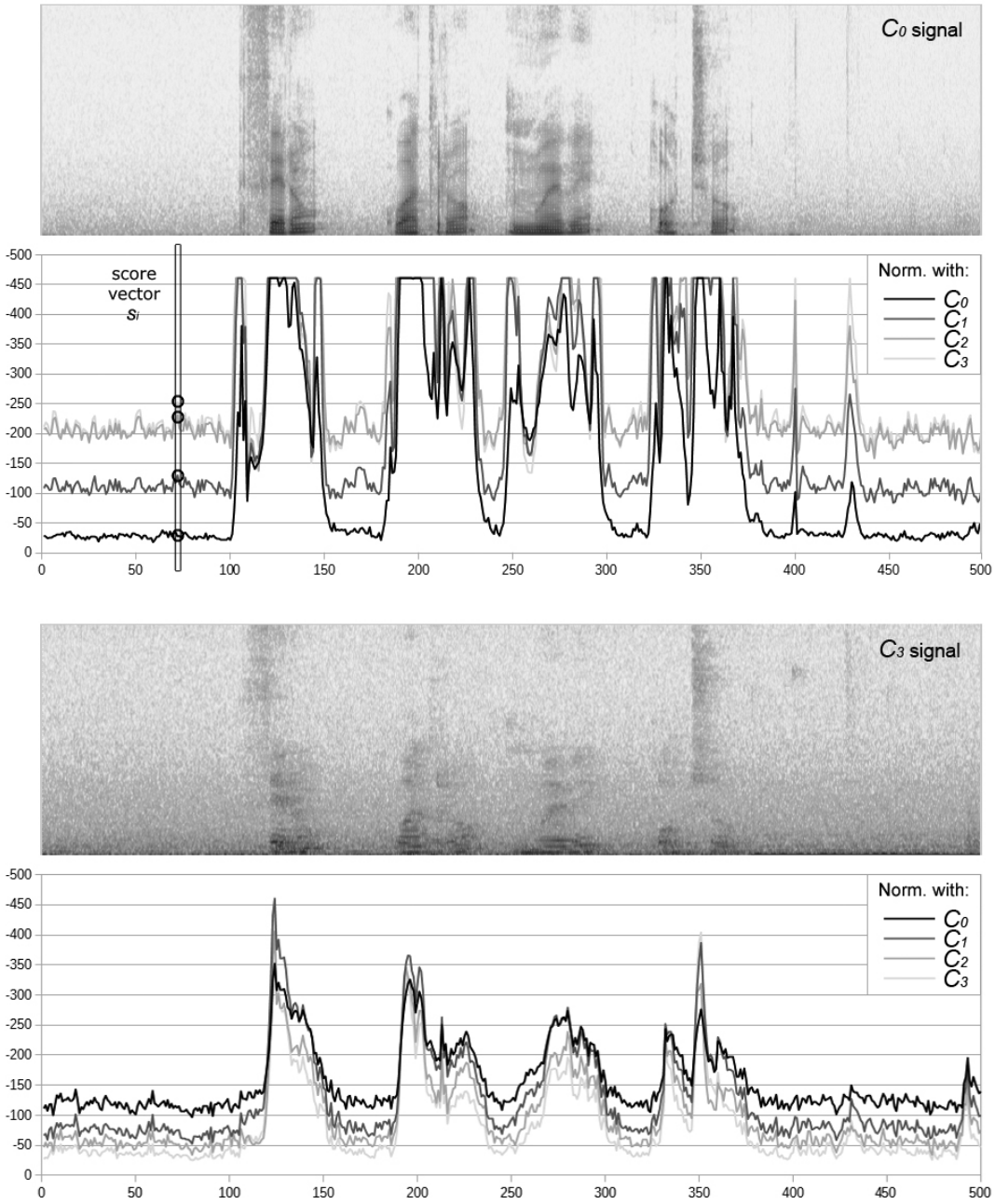


Figure 8.8: C_0 (goian) eta C_3 (behean) seinaleen espektrogramak eta isiltasun HMMaren erdiko egoerako behaketa-egiantzaren logaritmoak, denboran zehar (bilbeak), C_0, C_1, C_2 eta C_3 datu-multzoetatik aurrez kalkulatutako batezbestekoak eta bariantzak erabiliz hainbat modutan normalizatuta. Koadro bertikal estuak i bilbeko puntuazio-bektorea adierazten du.

8.4.1 Sailkatzailea: MLP

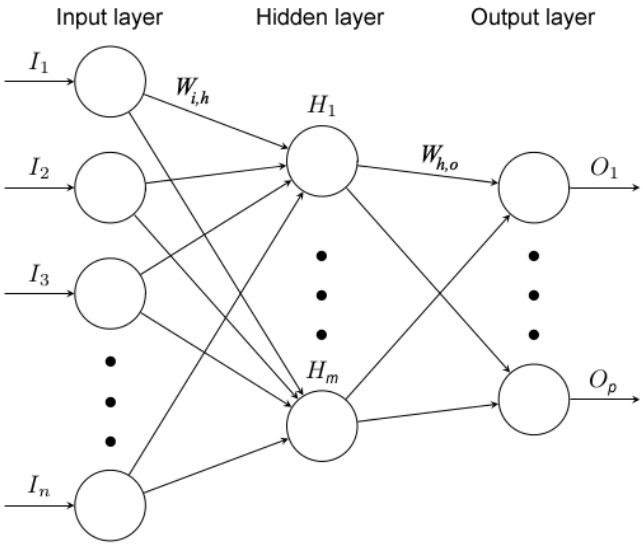


Figure 8.9: Ezkutuko geruza bakarreko MLP neurona-sarea, sarrerako geruzan, ezkutuko geruzan eta irteerako geruzan, hurrenez hurren, n, m eta p nodo dituena.

Sailkatzaile desberdinak testatu dira ikusteko ea modu desberdinetan normalizatutako parametro akustikoak erabiliz lortutako puntuazioak (normalizazio anitzeko puntuatzea edo MNS: *Multi Normalisation Scoring*) baliagarriak diren, eta emaitzarik onenak MLP batez lortu dira [188][189]. MLPa elikadura zuzeneko neurona-sare artifiziala da (ANN, *Artificial Neural Network*), sarrerako datuei irteera egokiak esleitzen dizkiena. Baldin eta sarrerako datuak behaketa berri baten atributu gisa eta irteerako datuak kategoria gisa kontsideratzen baditugu, prozesu hori sailkapen-atazatzat jo dezakegu. 8.9. irudian ageri denez, MLPa nodo-geruza anitzez osaturik dago grafo gidatu batean, eta geruza bakoitza hurrengoarekin osoro konektaturik dago. Sarrerako nodoak izan ezik, nodo bakoitza neurona (edo prozesaketa-elementu) bat da, aktibazio-funtzio ez-lineal bat duena. MLP eredu bat entrenatzeko, nodoen arteko pisuak kalkulatzen dira, errekurtsiboki, irteeraren eta espero den erantzunaren arteko desbiderapena neurtuz. Hori egiteko, atzeranzko hedapenaren teknika —ikasketa gainbegiratuko teknika bat— erabiltzen da [190]. Pertzeptroi lineal estandarrak [191][192] ez bezala, MLPak datu ez-linealki banangarriak bereizi ditzake [193]. MLPak, hurrengo esperimentuetarako, WEKA (*Waikato Environment for Knowledge Analysis*) erabiliz entrenatu dira, Java™ lengoaian idatzitako datu-meatzaritzako software libre eta ireki ezaguna [194, 195].

Table 8.3: MNSan oinarritutako VADa garatzeko erabilitako datu-baseen (eta kanalen) ezaugarri nagusiak.

Datu-basea	*Basque Speecon-like*		*Spanish Speecon - ECESS*				*TIMIT*	*Noisy TIMIT*	
Erabilitako kanalak	Hurbil.	Mahai gainekoa	Oso hurb. (C_0)	Hurb. (C_1)	Ert. (C_2)	Urr. (C_3)	Headset-mounted and far-field mic	*babble* noise (50-5 dB)	*white* noise (50-5 dB)
Hizkuntza	Euskara		Espainiera				Ingelesa (AEB)	Ingelesa (AEB)	
Ingurunea	Bulegoa		Bulegoa, leku publikoa, aisialdi-lekua, autoa				Estudioa	Estudioa + zarata gehigarria	
Hizlariak	230		60				630	630	
Fitxategiak / hizlari	316		17				10	600	
Eduki osoa (h)	109.95		1.41				5.37	322.2	
Hizketa-edukia (%)	47.90		51.77				86.57	86.57	
Etiketatzea	Lerrokatze fonetiko behartua		Eskuz				Eskuz	*TIMIT*etik	
Laginketa-maiztasuna	16 kHz		16 kHz				20 kHz (lagindua: 16 kHz)	16 kHz	

8.5 Hizketa datu-baseak

Atal honetan, MNSan oinarritutako VADaren esperimentuak egiteko erabili diren datubaseei buruzko informaziorik garrantzitsuena azaltzen da. Guztira, lau hizketa datu-base erabili ziren. Lehendabizi, *Basque Speecon-like* datu-basea [99] erabili genuen, hain zuzen hurbileko kanala, isiltasun-bilbeen HMMa entrenatzeko. HMM hori erabiliz, MLP bat entrenatu zen *Basque Speecon-like* datu-baseko fitxategiei eta ahots-aktibitatea eta ahostuntzea hautemateko ECESS ebaluazio-kanpainan erabilitako *Spanish Speecon* datu-basearen azpi-multzoko [180] fitxategiei MNS metodoa aplikatuz.

Lehen VAD esperimentua hirugarren datu-base bateko fitxategiak testatuz egin zen: *TIMIT Acoustic-Phonetic Continuous Speech* Corpusa [196]. Bigarren VAD esperimentua sistema seinale zaratatsuekin testatuz egin zen. Horretarako, *Noisy TIMIT* hizketa datu-basea [197] hautatu zen, hain zuzen ere murmurio zarata eta zarata zuriko datumultzoetako *Test* blokeak. Datu-multzo bakoitzak 10 azpimultzo ditu, eta azpimultzo bakoitza SNR desberdin bati dagokio (50 eta 5 dB bitartean, 5 dB-ko urratsetan). Hiru-

garren VAD esperimenturako, MLP berri bat entrenatu zen entrenamendu-materialari 10 azpimultzo horietako 4 (35, 25, 15 eta 5 *dB*) ere erantsita, sistema zaratarekiko sendoagotzeko asmoz. Testatutako fitxategiak bigarren esperimentuko berberak ziren.

Azkenik, emaitzak bi VAD algoritmo estandarrenekin alderatu ziren, 2. eta 3. esperimentuetako fitxategi berak testatuz: *Noisy TIMIT* hizketa datu-baseko *murmurio* zaratako eta zarata *zuri*ko datu-multzoetako *Test* blokeetakoak.

8.3. taulan, ikerketa honetan erabilitako datu-baseen eta datu-base bakoitzeko kanalen ezaugarri nagusiak ageri dira.

8.6 MNSan oinarritutako VADaren esperimentuak

Isiltasun HMMa Basque Speecon-like datu-basearen Train blokea erabiliz entrenatu zen. Audio-seinaleak 25 *ms*-ko luzerako bilbez leihokatu ziren, bilbe bat jasoz 10 *ms*-ko. Parametro akustikoak 13 MFCCz eta lehen mailako 13 eta bigarren mailako 13 deribatuz osaturik daude, eta 32 gaussiarreko GMMz modelatu ziren. Parametro akustiko horiek normalizatu egin ziren, saio bereko fitxategietatik (hizlari berari dagozkion fitxategietatik, alegia) erauzitako batezbestekoak eta bariantzak erabiliz.

8.6.1 MNSan oinarritutako VAD esperimentua, MLP bat erabiliz

Entrenamendu-datuak prestatzeko, sei datu-multzo baliatu ziren: *Spanish SpeeCon* datu-baseko C_0, C_1, C_2 eta C_3, eta *Basque Speecon-like* data-baseko *hurbileko* eta *mahai gaineko* kanalak. Datu-multzo guztiek 1 020 fitxategi dituzte, eta datu-base berekoek hizketa-eduki bera dute. Horrek esan nahi du berdina dela, adibidez, C_0 datu-multzoko 1 020 fitxategien edukia eta C_3 datu-multzoko fitxategiena; eta halaber berdinak direla *hurbileko* datu-multzoko 1 020 fitxategien edukia eta *mahai gaineko* datu-multzoko fitxategiena. Datu-multzo bakoitzeko fitxategiak prozesatuta, parametro akustikoak normalizatu egin dira, haietatik kalkulatutako batezbestekoak eta bariantzak erabiliz, eta, segidan, behaketa-egiantzen logaritmoak erauzi dira. Hala, 6 puntuazioko bektoreak lortzen dira fitxategi bakoitzeko bilbe bakoitzerako. Guztira, 3 096 632 puntuazio-bektore erauzi dira, haietariko % 49.08 hizketa izanik eta % 50.92, berriz, isiltasuna. Horrek erakusten du ondo orekaturik dagoela entrenamendurako erabili dugun datu multzoa. 8.4. taulan datuei buruzko informazio gehiago ageri da.

Ataza honetarako erabilitako MLPak 6 nodo ditu sarrerako geruzan (bat puntuazio bakoitzeko) eta 2 nodo irteerako geruzan (bat kategoria bakoitzeko). Bi nodo kopuruen baturaren erdia (4 nodo) hautatu da ezkutuko geruzarako.

Spanish SpeeCon datu-baseko fitxategien kasuan, datu-baseko etiketak erabili dira, entrenamenduko bilbe bakoitzari "hizketa" edo "isiltasuna" kategoria esleitzeko. Etiketa horiek "sil" (isiltasuna edo ez-hizketa), "u" (*unvoiced speech*, hizketa ahoskabea) eta "v" (*voiced speech*, hizketa ahostuna) gisa etiketatutako segmentuetatik erauziak dira. Bistan da "u" eta "v" segmentuak erabili direla "hizketa" kategoriarako. *Basque Speecon-like* datu-basearen kasurako, HMMaren entrenamendu-prozesutik atera dira etiketak, fonema-lerrokatzeen bitarteko fitxategietatik.

Table 8.4: MLParen entrenamendu-datuak osatzeko erabilitako datu-multzoak, fitxategi kopuruak eta bilbe kopuruak.

	Basque Speecon-like db	*Spanish SpeeCon* db	
Datu-multzoak	*hurbilekoa*, *mahai gainekoa*	C_0, C_1, C_2, C_3	GUZTIRA
Fitxategiak	1 020 × 2	1 020 × 4	6 120
isiltasun bilbeak	299 972 × 2	244 211 × 4	1 576 788
hizketa bilbeak	234 628 × 2	262 647 × 4	1 519 844
Bilbeak guztira	534 600 × 2	506 858 × 4	3 096 632

MNSan oinarritutako MLParen hasierako testa egiteko, beste datu-base bat hautatu da: *TIMIT Acoustic-Phonetic Continuous Speech* Corpusa [196]. Guztira, 6 300 esakuntzaz osaturik dago, Amerikako Estatu Batuetako 8 dialekto nagusietako 630 hizlarik esanak, bakoitzak 10 esaldi. 10 esaldi horietan, 30 segundoko hizketa-edukia dago, batez bestez, hizlari bakoitzeko. Guztira, corpusak 5 hizketa-ordu inguru ditu. Arrazoi praktikoak direla eta, zenbait eremu dialektal ez daude gainerakoak bezain ondo errepresentaturik. Gauza bera gertatzen da emakumezko-hizlariekin; izan ere, gaizki orekatuta daude gizonezko-hizlariekin alderatuta: emakumezkoak, % 30; gizonezkoak, % 70.

Buruko batean muntatutako Sennheiser HMD 414 mikrofonoa eta urrutiko Breul & Kjaer 1/2" presiozko mikrofonoa erabili ziren, bi kanaleko seinaleak grabatzeko. Hizketa zuzenean digitalizatu zen, 20 kHz-eko laginketa-maiztasunean, *aliasing*aren kontrako 10 kHz-eko iragazki batez, eta 16 kHz-etara jaitsi zen gero, laginketa-maiztasuna. Dena dela, Sennheiser mikrofonoaz grabatutako hizketa-datuak soilik ezarri ziren CD-ROMaren banaketa-bertsioan, eta hori da guk esperimentu honetan erabili duguna.

TIMIT erabiltzearen abantaila nagusia da zehatz-mehatz etiketatuta dagoela: "h#" etiketatutako segmentuek hasierako/amaierako isilunea adierazten dute; "pau" etiketek, berriz, tarteko isiluneak. Hortaz, bi etiketa horiez markatutako segmentuak hartu dira kontuan, "isitasun" edo "ez-hizketa" gisa; gainerako segmentuak "hizketa" gisa etiketatu dira. Guztira, 15 ms-ko luzerako bilbeak kontsideratuz, 10 ms-ko bilbe-maiztasunaz jasoak, 1 925 077 bilbe daude *TIMIT* datu-basean. Haietatik % 86.57 "hizketa" bilbeak dira, eta % 13.43 soilik "isiltasun" bilbeak.

Bi test egin dira esperimentu honetan: batetik, *TIMIT* fitxategi guztiak (6 300) prozesatu dira, *offline* moduan; alegia, fitxategi bakoitzeko batezbestekoak eta bariantzak aurrez kalkulatuta. Orduan, parametro akustikoak normalizatzen dira, eta aurreko atalean kalkulatutako atalasea (Th = -150) erabiltzen da bilbe bakoitza "hizketa" edo "isiltasun" gisa sailkatzeko. Bestetik, *TIMIT*eko fitxategi berak prozesatu dira *online* moduan; hau da, parametro akustikoak kalkulatu ahala normalizatzen dira, 6 modutara, 6 puntuazioko bektore bat lortuz. Lortu ahala sailkatzen dira puntuazio horiek, MLParen bidez. *Offline*eko eta *online*eko esperimentuen emaitzak 8.5. taulan ageri dira.

Table 8.5: *TIMIT* corpusaz egindako *offline*ko eta *online*ko VADaren esperimentuetako TER, ER_0 eta ER_1.

	TER	ER_0	ER_1
Offline esperimentua (Th = -150)	5.27	32.67	1.03
Online esperimentua (MLP)	4.98	19.68	2.70

Emaitzek erakusten dute MNSan oinarritutako MLPa erabiliz emaitza pittin bat hobeak lortzen direla. Izan ere, TERari dagokionez, % 5.50eko hobekuntza lortu da *offline* esperimentuarekin alderatuta. ER_0 ere nabarmen hobetu da. Hala ere, ER_1 emaitzak okerragotu egin dira, nahiz eta, halere, baxuak izan.

Kontuan izan behar da *offline* esperimentuko atalasea *Basque Speecon-like* datu-basea soilik erabiliz kalkulatu dela. Beraz, *Spanish SpeeCon* datu-baseko C_0 azpimultzoa testatzeko esperimentuen emaitzak (8.1. taula, C_0 lerroa) eta *TIMIT* datu-basekoak (8.5. taula) baliokideak dira, atalasea kalkulatzeko datu-multzoa ez beste bat testatzen dute eta. Emaitzak antzekoak dira, ER_0rako izan ezik, *TIMIT* datu-baseko fitxategiak testatuz baino balio altuagoak lortzen baitira. Horren zergatia, seguruenik, datu-baseen arteko desberdintasun akustikoetan datza. Hala eta guztiz ere, baxu jarraitzen du ratio garrantzitsuenak (ER_1), eta are baxuagoa da *TIMIT* testean.

Online esperimentuari dagokionez, *Basque Speecon-like* eta *Spanish SpeeCon* datu-baseetako datuak erabiliz entrenatu da MLPa. Baliteke, hortaz, hori izatea esperimentu honetan emaitza orokor hobeak lortzearen arrazoia, alegia, iturri gehiagoko datuak erabili izana. Dena dela, MNS teknika atalasea kalkulatzeko ideiaren hedapen landua da. Horrek ere azaldu dezake hobekuntza.

Laburbilduz, atal honetan azaldutako esperimentuak erakusten du ezen lan honetan proposatutako MNS teknika, gutxienez, *offline* teknika bezain eraginkorra dela. MNS teknikaren abantaila nagusia da 15 ms-ko luzerako audio-bilbe bat sailkatzeko gai dela, eredu bat erabiliz; hau da, inguruko bilbeak edo uneko bilbea dagoen segmentu (edo fitxategi) bateko bilbeak analizatu beharrik gabe. Ideia hori itxaropentsua da, zeren hura inplementatzeak ez luke atzerapenik eragingo, eta oso azkarra izango litzateke. Hala eta guztiz ere, atal honetan deskribatutako esperimentuan lortutako emaitzak seinale garbiak testatuz (*TIMIT* datu-baseko fitxategiak) atera dira, eta, horrenbestez, hizketa zaratatsua testatu behar da, teknika berri horren sendotasuna eta benetako erabilgarritasuna balioesteko. Hurrengo atalean azaltzen da hori.

8.6.2 MNSan oinarritutako MLP esperimentuak baldintza zaratatsuetan

Aurreko atalean aurkeztutako teknikaren sendotasuna balioesteko, hizketa zaratatsua duten fitxategiak testatuz ebaluatu behar da. Horretarako, Floridako Teknologia Institutuan garatutako datu-base bat hautatu da: *Noisy TIMIT* hizketa datu-basea [197]. Datu-base horrek 322 ordu inguruko hizketa du, TIMIT datu-baseko fitxategiei [196] zarata gehigarri desberdinak eta neurri desberdinetan erantsiz egina. Hortaz, haren

etiketa-fitxategiak *TIMIT* datu-base klasikoari dagozkionen berdinak dira.

Zarata gehigarriak hauek dira: *murmurioa*, *zuria*, *arrosa*, *urdina*, *gorria* eta *bioleta*; eta zarata mailak 50 dB-tik 5 dB-ra bitartekoak dira, 5 dB-ko mailatan. Zarata-koloreak zarata-seinalearen espektro-dentsitateari egiten dio erreferentzia. Zarata koloretsuak argiaren koloreen modu berean izendatzen dira. Adibidez, zarata *zuriak* maiztasun entzungarri guztiak ditu, alegia, argi *zuriak* eremu ikusgarriko maiztasun guztiak dituen era berean. Hala, zuriak ez diren koloretako zaratek energia gehiago dute kontzentratuta soinu-espektroaren goiko edo beheko muturrean. Audio-fitxategi guztiak kanal bakarreko seinale gisa daude gordeta, 16 kHz-etan eta 16-flac formatuan, baina 16-bit PCM (WAV) formatura bihurtu dira.

Esperimentuen funtsa da aurreko atalean erabilitako MLP bera erabiltzea, baina *Noisy TIMIT* datu-baseko fitxategi guztiak testatuz. Hemen ere, 15 bilbe hartu da bai hizketa-segmentuen, bai isiltasun-segmentuen gutxieneko iraupen gisa. Zarata mota bakoitzerako eta hainbat SNRtarako lortutako TERak 8.10. irudian ageri dira, non *TIMIT*eko erreferentzia (8.5. taula) lerro eten gisa adierazirik baitago.

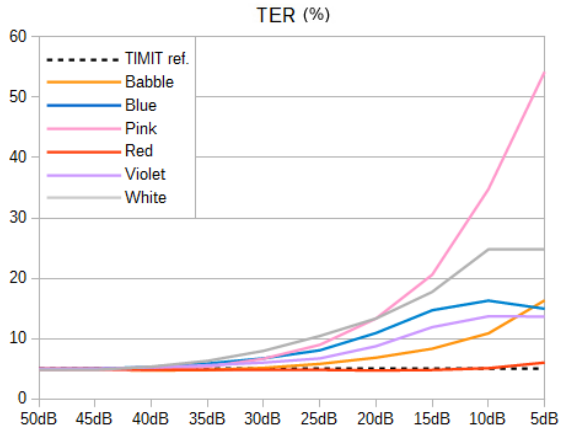


Figure 8.10: *Noisy TIMIT* datu-basea MNS teknikaz testatuz lortutako VAD TERak, hainbat zarata motatarako (koloreak) eta SNR mailatarako.

Emaitzek adierazten dute zarata mota bakoitzak eragin desberdina duela VADaren funtzionamenduan. Lehen begiratuan, zarata *gorria* (espektroaren beheko muturrerantz nabarmen lerratua) da kasurik onena, ia ez baitu eraginik emaitza orokorretan; aitzitik, zarata *zuria* (maiztasun-espektro laukoa) eta zarata *arrosa* (proportzionalki txikiagotzen dena maiztasuna handiagotu ahala) dira kaltegarrienak. Bigarren eraginik txikiena duena *murmurio* zarata da (5 dB-ko kasurako izan ezik). Hala ere, aztertutako hizketa- eta isiltasun-bilbeen kopuruen arteko aldea dela eta (hizketa-bilbeak kopuru osoaren % 86.57 dira), emaitza orokorrak ER_1 emaitzen oso antzekoak dira. ER_0 eta ER_1i dagokienez, 8.11. irudian ageri dira emaitzak, non TIMITeko erreferentziak (8.5. taula) lerro eten gisa adierazirik baitaude.

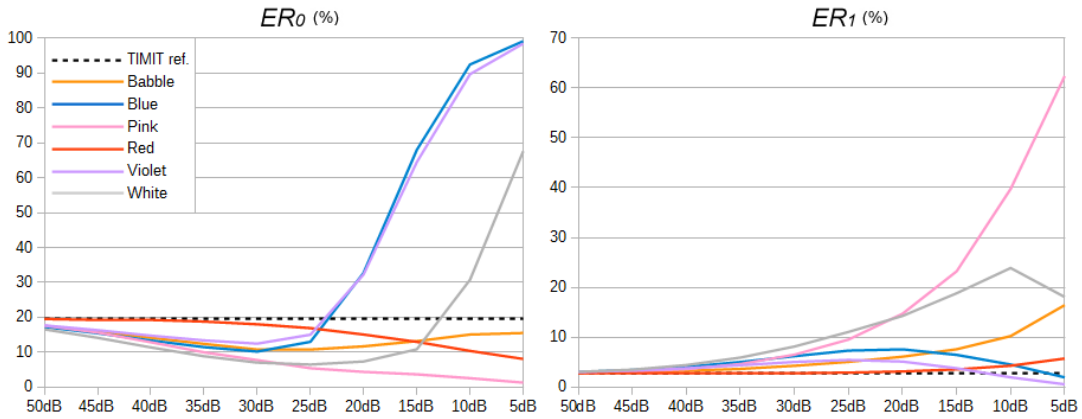


Figure 8.11: *Noisy TIMIT* datu-basea MNS teknikaz testatuz lortutako VAD ER_0ak eta ER_1ak, hainbat zarata motatarako (koloreak) eta SNR mailatarako.

ER_0 eta ER_1 emaitzak batera aztertzen baditugu, ondoriozta dezakegun zarata *gorria* dela seguruena. Zarata *urdina* (proportzionalki handiagotzen dena maiztasuna handiagotu ahala) eta zarata *bioleta* (espektroaren goiko muturrerantz nabarmen lerratua) hartuta, zenbat eta SNR baxuagoa, orduan eta isiltasun-bilbe gehiago etiketatuko dira hizketa gisa. Horrek esan nahi du puntu bat iritsiko dela non bilbe guztiak hizketa gisa etiketatuko baitira; hori VADrik ez erabiltzea bezala da. Kontrako kasua zarata *arrosa*ren kasuan gertatzen da. Zenbat eta SNR baxuagoa, orduan eta hizketa-bilbe gehiago etiketatuko dira isiltasun gisa. VADarentzako kasurik txarrena da hori, zeren hizketa-bilbe gehiago eta gehiago etiketatuko dira isiltasun gisa, eta baztertu egingo dira, beraz. Zarata *zuri*ari dagokionez, bi kasuetako ezaugarriak ageri dira. Horrek esan nahi du kasu horretan gorabeheratsuagoa dela sailkapena.

Murmurio zaratari dagokionez, espero izatekoa litzateke, hasiera batean, ezen zenbat eta 0 dB-tik hurbilago SNRa, orduan eta isiltasun-bilbe gehiago hizketa gisa sailkatuak. Hala ere, emaitzek bestelako jokaera bat azaltzen dute: ER_0ren kurbak 25 dB-an % 10.75eko minimo bat du, eta % 15.55 balio du 5 dB-an (50 dB-an: % 17.35); ER_1, berriz, handiagotu egiten da SNR txikiagotu ahala. Isiltasun-bilbeak gero eta okerrago sailkatzen dira SNR 0 dB-ra hurbildu ahala, eta hori ez da onargarria VAD sistema baterako.

Laburbilduz, baldintza zaratatsuetan egindako MNS esperimentuek MLP sailkatzailearen oso jokaera gorabeheratsua erakusten dute, zarata motaren arabera. Dena dela, kontuan izan behar da MLPa entrenatzeko erabilitako fitxategiak ez direla esperimentu honetan testatutakoak bezain zaratatsuak. Horrenbestez, beste esperimentu bat egitea pentsatu da, ikusteko ea MNSan oinarritutako MLPa gai den baldintza zaratatsuetan hizketa eta isiltasuna (edo ez-hizketa) modelatzeko.

8.6.3 MNSan oinarritutako MLPari seinale zaratatsuak erantsiz

Esperimentu berri honetan, seinale zaratatsuak erantsi zaizkio MLParen entrenamendu-prozesuari. Murmurio-zarata eta zarata zuria soilik hartu dira kontuan; izan ere, bi horiexek kontsidera ditzakegu naturalenak, tesi honetan proposatutakoaren moduko sistema baterako. Hortaz, *Noisy TIMIT* datu-baseko murmurio-zarataren eta zarata zuriaren azpimultzoak gehitu zaizkie aurreko esperimentuan MLPa entrenatzeko erabilitako *Basque Speecon-like* eta *Spanish SpeeCon* datu-baseetako fitxategiei. Fitxategi zaratatsu horiek azpimultzo bakoitzeko *Train* blokekoak dira, eta azpimultzo bakoitzean 10 SNR desberdinetariko 4 seinale multzo erabili dira: 50, 35, 20 eta 5 dB. Horrek erakutsiko digu ea nola dabilen sistema datu zaratatsuen multzo txiki bat erabiliz.

Train blokea aztertuz, ikusi da alde handia dagoela isiltasun- eta hizketa-bilbeen kopuruen artean: 190 052 isiltasun-bilbe and 1 220 017 hizketa-bilbe. *Spanish SpeeCon* datu-basetik hartutako isiltasun-bilbeen kopurua 244 211 da; *Basque Speecon-like* datu-basekoak, berriz, 299 972 (ikus 8.4. taula). Hortaz, nahiz eta pittin bat desorekaturik egon, *Train* blokeko fitxategi guztiak erabili dira isiltasun-bilbeen gehienezko kopuruari eusteko. Bestalde, hizketa-bilbeen kopurua 244 003ra murriztu da ausaz. Kopuru hori 8.4. taulan erabilitako kopuruen antzekoa da. Guztira, *Noisy TIMIT*eko 1 520 416 isiltasun-bilbe eta 1 952 024 hizketa-bilbe erabili dira, bilbe kopuru osoa 6 569 072 izanik (3 097 204 isiltasun-bilbe, 3 471 868 hizketa-bilbe). 8.6. taulak informazio hori guztia du.

Table 8.6: MLParen entrenamendu-datuak seinale zaratatsuz eraikitzeko baliatutako datu-mulzoak, fitxategi kopuruak eta bilbe kopuruak.

	Basque Speecon-like	*Spanish SpeeCon*	*Noisy TIMIT*	
Datu-multzoak	*hurbilekoa*, *mahai gainekoa*	C_0, C_1, C_2, C_3	*murmurioa*, *zuria*: 50, 35, 20, 5 dB	GUZTIRA
isilt. bilbeak	299 972 × 2	244 211 × 4	190 052 × 8	3 097 204
hizketa bilbeak	234 628 × 2	262 647 × 4	244 003 × 8	3 471 868
Bilbeak guztira	534 600 × 2	506 858 × 4	434 055 × 8	6 569 072

Puntuazio-bektoreek 14 elementu dituzte orain: 2 puntuazio *Basque Speecon-like* datu-baseko hurbileko eta mahai-gaineko azpimultzokoak dira; 4 puntuazio *Spanish SpeeCon* datu-baseko *C0*, *C1*, *C2* eta *C3* azpimultzoetakoak; eta 8 puntuazio *Noisy TIMIT* datu-baseko murmurio-zarataren eta zarata zuriaren azpimultzoetatik hautatutako 4 SNR maila desberdinetakoak (50, 35, 20 eta 5 dB). 14ko tamainako puntuazio-bektore horiekin, esperimenturako hautatu den MLP konfigurazio honako hau da: 14 nodo sarrerako geruzan, 2 nodo irteerako geruzan, eta bi nodo kopuruen baturaren erdia (8 nodo) ezkutuko geruzan.

Testatutako fitxategiak 10 080 dira, murmurio-zarataren eta zarata zuriaren azpimultzoetako *Test* blokeei dagozkienak, SNR maila erabilgarri guztietan (1 260 fitxategi SNR talde bakoitzean). Emaitzak (TER) 8.12. irudian ageri dira (lerro jarraituak), aurreko esperimentuan lortutako TERekin batera (lerro etenak). Lehen begiratuan, ematen du murmurio-zarataren kurbak hobekuntza txikiagoa duela; zarata zuriak, ordea, hobekuntza nabarmena du, batez ere SNRa txikiagoa egin ahala. 5 dB-an, murmurio-zarataren errore-tasa % 12.15 da, eta % 25.69ko hobekuntza du; zarata zuriaren errore-tasa, berriz, % 8.44 da, eta % 65.87ko hobekuntzako du. Azpimarratzekoa da bi kurbek orain profil antzekoagoa osatzen dutela.

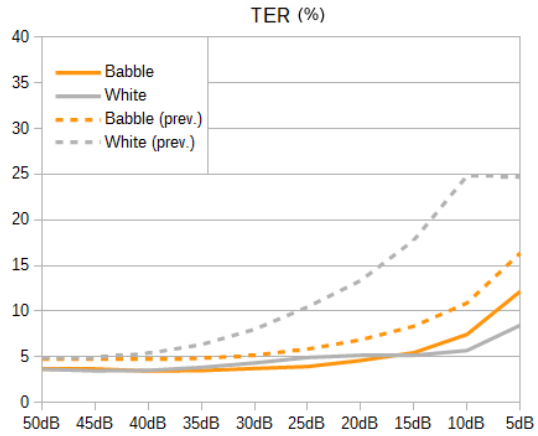


Figure 8.12: Noisy TIMIT datu-basearen murmurio-zarataren eta zarata zuriaren azpimultzoetako *Test* blokeak testatuz lortutako VAD TERak, SNR maila erabilgarri guztietarako (lerro etenak: aurreko esperimentuko emaitzak).

Noisy TIMIT datu-baseko hizketa- eta isiltasun-bilbeen kopuruen arteko aldearen ondorioz, ER_0 eta ER_1 aztertu behar dira, interpretazioa fidagarriagoa eta errealistagoa izan dadin. ER_0 eta ER_1 balioak 8.13. irudian ageri dira. ER_1i dagokionez, TER emaitzen nahiko antzekoak dira emaitzak, lehen azaldutako desoreka dela eta. Hortaz, azalpen bera eman diezaiokegu, are gehiago kontsideratzen badugu emaitzak hobetu egin dira baldintzarik garbienetan. 5 dB-an, % 11.09ko ER_1 lortu da murmurio-zaratarako, eta % 7.52, zarata zurirako. Bestalde, ER_0ri dagokionez, emaitzek eragin desberdina erakusten dute murmurio-zaratako eta zarata zuriko seinaleetan: murmurio-zaratarako kurba aurrekoaren antzekoa da, baina orain balioak pittin bat txarragoak dira, SNR balio ertainetarako izan ezik (25 dB ingururako). Zarata zuriaren azpimultzoari dagokionez, ER_0 kurbak ere emaitza txarragoak ditu atal garbienerako, baina askoz emaitza hobeak baldintza zaratatsuenetarako, 15 dB-tik aurrera izugarri igotzen baita. Gainera, emaitza hobea da 5 dB-an (% 14.51), 50 dB-an (% 18.26) baino. Baiezta liteke, beraz, entrenamendu-prozesuan datu zaratatsuak erabilita, egonkortu egiten direla datu zaratatsuak testatzearen emaitzak.

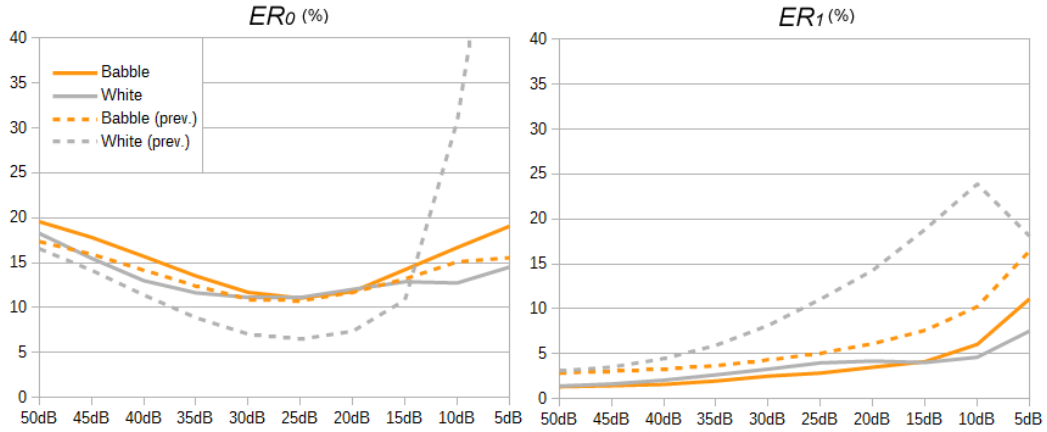


Figure 8.13: *Noisy TIMIT* datu-baseko murmurio-zarataren eta zarata zuriaren azpimultzoen *Test* blokeak MNS teknikaz testatuz lortutako VAD ER_0ak eta ER_1ak, SNR maila erabilgarri guztietarako (lerro etenak: aurreko esperimentuko emaitzak).

Merezi du azpimarratzea % 50eko hobekuntza lortzen dela seinale garbienekin (50 dB), eta 8.5. taulan lortutako *offline* emaitzatik hurbilago geratzen da. Hori oso emaitza lehiakorra da (ikus 8.7. taula).

Table 8.7: *TIMIT* corpuseko *online* VAD esperimentuko TER, ER_0 eta ER_1ak.

		TER	ER_0	ER_1
Offline esperimentua (Th = -150)		5.27	32.67	1.03
Online esperimentua		4.98	19.68	2.70
Online esp. (+ datu zaratatsuak)	murmurio-zarata 50 dB	3.73	19.57	1.32
	zarata zuria 50 dB	3.63	18.26	1.40

Oro har, emaitzek adierazten dute ezen, entrenamendu-prozesuan datu zaratatsuak sartuta, antzeko ER_1 kurbak lortzen direla murmurio-zaratarako eta zarata zurirako. Bestalde, ER_0 emaitzak okerragotu egiten dira, pittin bat, seinale garbienetarako; hobetu egiten dira, ordea, zaratatsuenetarako (nabarmen hobetu, gainera, zarata zurirako).

8.6.4 Beste zarata mota batzuetara orokortzea

Ikusi dugu murmurio datu-multzoko eta zuri datu-multzoko 4na azpimultzo (35, 25, 15 eta 5 dB) erabiliz entrenatutako MLPa gai dela emaitzak gainerako SNR balioetarako orokortzeko. Hala ere, ikusi nahian ea seinale zaratatsuz entrenatutako MLPa (ikus 8.6.3. atala) beste zarata mota batzuetarako ere baliagarria izan daitekeen, beste zarata batzuk dituzten seinaleekin testatu genuen MLPa. Hala, test-multzorako, orain,

murmurio eta zuri datu-multzoetako *Test* blokeetako fitxategiez gainera, urdin, arrosa, gorri eta bioleta datu-multzoetako *Test* bloke beretako fitxategiak ere erabili ziren (1260 fitxategi SNR azpimultzo bakoitzean; 12 600, datu-multzo bakoitzean).

8.14. irudian, zarata mota desberdinetarako hainbat SNRtan lortutako TERak ageri dira. 35 dB-tarako edo hura baino SNR handiagoetarako, ez dago degradaziorik entrenatzeko erabili ez diren zaratak dituzten seinaleak testatzean. SNR txikiagoetarako, okerragotzea ez da oso handia: handiena zarata bioletarako da, 15 dB-tan bi erreferentziekiko 7 puntu okerragotzen baita. Zarata gorrirako, gainera, erreferentziazko zaratak testatzen direnean baino hobea da sistemaren funtzionamendua.

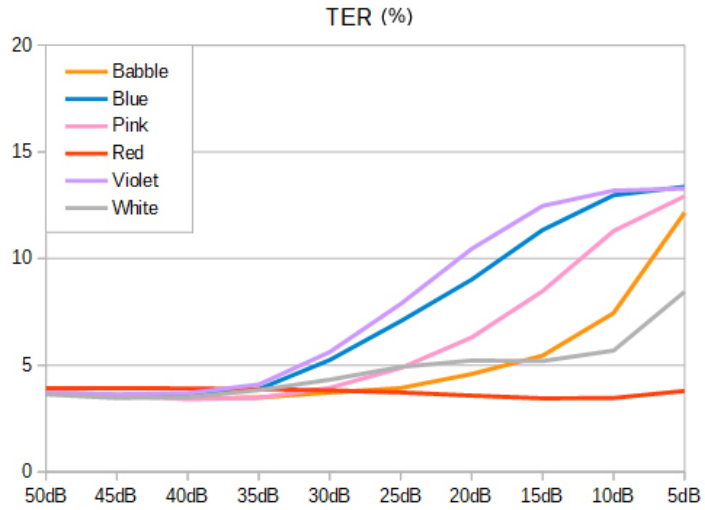


Figure 8.14: Murmurio-zarata eta zarata zuria duten seinaleekin entrenatutako MLPa erabiliz *Noisy TIMIT*eko zarata mota guztietako SNR guztiak testatuz lortutako TERak

8.7 Azken esperimentuak

ITU-T (*International Telecommunication Union - Telecommunication Standardization Sector*) erakundeak estandarizatutako bi *online* VAD algoritmo testatu ziren, guk proposatutako VAD teknikaren baliagarritasuna egiaztatzeko. Algoritmoak G seriekoak dira (*Transmission systems and media, digital systems and networks*), non *G.710 - G.729* bitartekoak "ahots- eta audio-seinalearen kodeketa"ri baitagozkio. Lehen algoritmoa *G.720.1* da [198]; berez, soinu-aktibitatearen detektatzaile generikoa (GSAD, *Generic Sound Activity Detector*) da, 8 edo 16 kHz-eko audioa prozesatzen duena, eta VAD modulu bat du. Bigarren modulua *G.729* da [199], 8 kHz-eko sarrerako seinaleak onartzen dituen 8 kbit/s-ko hizketa-kodetzailea, *G.729*ko B eranskinean deskribatutako VAD moduluan oinarrituta dagoena (*G.729b* izenez ere ezagutzen da). Sistema biek erabiltzen dute 10 ms-ko bilbe-luzera eta bilbe-desplazamendua, eta ez dute *look-*

ahead edo aurrera begiratzerik behar (ez da atzerapenik sortzen, beraz; soilik bilbearen iraupena). Xehetasun gehiago ageri dira 8.8. taulan, bi ITU sistemetarako[1] eta guk proposatutako VAD teknikarako.

Kontuan izan behar da konputazio-denbora dela 10 080 fitxategiko test batean sistema bakoitzak fitxategiko behar duen batez besteko denbora, ordenagailu bera erabiliz eta baldintza beretan. Litekeena da desberdina izatea ordenagailu batetik bestera, baina sistemen arteko ratioei antzematen laguntzen digu. Bestalde, suspertze-eskemari dagokionez, *G.729b* algoritmoak eta guk proposatutako VAD teknikak antzeko egoera-makina dute, eta atzerapen labur bat sortzen dute, egoera-aldaketa dagoen ala ez erabaki bitartean. *G.720.1* algoritmoaren kasuan, eskema kontserbadoreago jarraitzen du, non adierazle aktiboak igortzen baitira isiltasun-segmentu bat detektatzen den arte.

Table 8.8: *G.720.1* algoritmoko (ITU-T), *G.729b* algoritmoko (ITU-T) eta guk proposatutako VAD teknikako zenbait parametro garrantzitsuren konparaketa.

	G.720.1 VAD	*G.729b*	Prop. metodoa
Banda-zabalera (kHz)	8, 16	8	16
Bilbe-iraupena / -desplazamendua (ms)	10 / 10	10 / 10	25 / 10
Konputazio-denbora (ms fitxategiko)	26.8	34.87	30.7
Leuntzea	Ez	Bai	Bai
Hasieratzea (bilbe kopurua)	200 ez-aktibo	32	0

VADak testatzeko, 8.6.2. ataleko eta 8.6.3. ataleko datu berak erabili ziren. *G.729b* kodetzailean oinarritutako VADa testatzeko, fitxategien maiztasuna 8 kHz-etara iragan zen. 8.15. irudian, isiltasun-sailkapenaren errore-tasak (ER_0) ageri dira, bi ITU algoritmoak (lerro jarraituak) eta guk proposatutako metodoa (lerro etena) erabiliz lortutakoak, bai murmurio-zaratadun, bai zarata zuridun seinaleetarako. Bi ITU sistemen minimoak % 30etik gora daude, eta nabarmen handiagotzen dira SNRa txikiagotu ahala, batez ere murmurio-zarata duten seinaleetan. Horrek esan nahi du ezen, seinalea zenbat eta zaratatsuago, orduan eta isiltasun-bilbe gehiago klasifikatuko direla hizketa gisa.

Bestalde, 8.16. irudiak hizketa-sailkapenaren errore-tasak (ER_1) erakusten ditu. *G.720.1*en kasuan, murmurio-zaratadun seinaleen emaitzak eta guk proposatutako metodoaren emaitzak oso antzekoak dira. Zarata zuriari dagokionez, emaitzak ere berdinak dira zarata baxuko seinaleetarako, baina guk proposatutako metodoak emaitza hobeak ditu datu zaratatsuenekin. *G.729b*-ri dagokionez, guk proposatutako metodoak emaitza hobeak ditu 25 dB baino SNR handiagoetarako; hortik behera, *G.729b* algoritmoak funtzionamendu hobea du. 10 eta 5 dB-tan, *G.729b*-k emaitza hobeak ditu

1 The software for both systems can be downloaded from the ITU website: `http://www.itu.int/rec/T-REC-G.720.1-201001-I` and `http://www.itu.int/rec/T-REC-G.729-201206-I`, respectively.

murmurio-zaratarako; zarata zurirako, berriz, MNSan oinarritutako VADak.

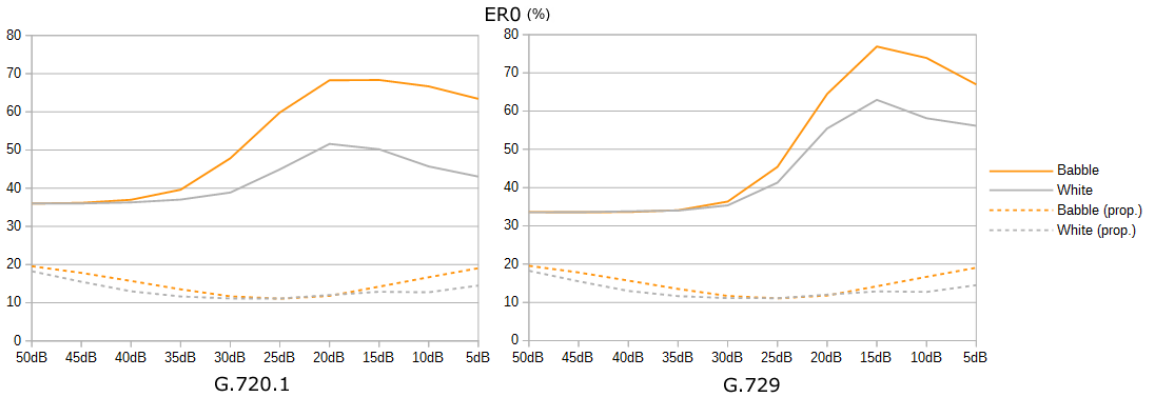


Figure 8.15: *Noisy TIMIT*eko murmurio-zaratadun eta zarata zuridun seinaleekin ITU-T *G.720.1* (ezkerrean) eta *G.729b* (eskuinean) VAD estandarrak erabiliz lortutako VAD ER_0ak, guk proposatutako sistemaren emaitzekin batera (lerro etena).

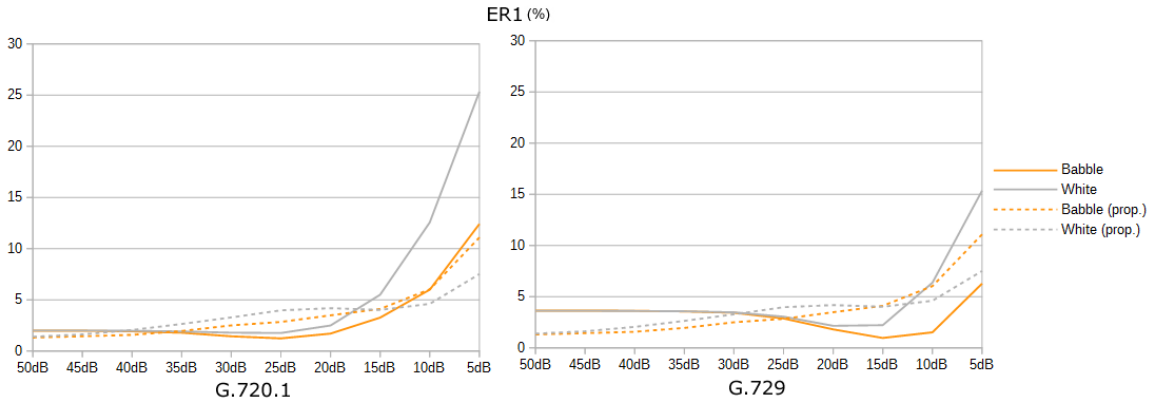


Figure 8.16: Noisy TIMITeko murmurio-zaratadun eta zarata zuridun seinaleekin ITU-T *G.720.1* (ezkerrean) eta *G.729b* (eskuinean) VAD estandarrak erabiliz lortutako VAD ER_1ak, guk proposatutako sistemaren emaitzekin batera (lerro etena).

Oro har, ITU algoritmoekin eta MNSan oinarritutako VADarekin lortutako ER_1 emaitzak alderagarriak dira SNR altuko seinaleentzat. Eta SNR baxuko seinaleentzat, alderantziz, emaitzek jokaera desberdina dute murmurio zaratarako eta zarata zurirako. Murmurio-zaratarako, *G.720.1*ek antzeko emaitzak ditu; *G.729b*-k, berriz, emaitza hobeak. Zarata zurirako, emaitza hobeak lortzen dira MNSan oinarritutako sistemaz. Hala ere, azpimarratzekoa da ER_0 balioak oso altuak direla bi ITU algoritmoentzat; horrek esan nahi du ezen, seinale zaratatsuak testatzean, bi sistemek dutela isiltasun-

bilbeak hizketa gisa sailkatzeko joera.

Ondorioz, guk proposatutako VAD teknikak TER hobeak lortzen ditu zarata-maila guztietan. Hizketa-bilbeen eta isiltasun-bilbeen kopuruen arteko kopuruen arteko desorekaren ondorioz, TER kurbak ER_1 kurben antzekoak dira, baina ER_0z proportzionalki desplazatuak. ITU sistemen abantailetariko bat da zarata-baldintza desberdinetara egokitu daitezkeela martxan dauden bitartean; hala ere, hasieraketa-denbora bat behar dute parametro nagusiak doitzeko. MNSan oinarritutako gure sistema, ordea, entrenamenduan erabili ez diren zarata motetara orokortzeko gauza da, eta ez du hasieraketa-denborarik behar, emaitzak ez baitira aurreko bilbeen mendekoak.

8.8 Konklusioak

Kapitulu honetan, VAD teknika berri bat deskribatu da: isiltasun HMMaren erdiko egoerako GMMak sortutako behaketa-puntuazioen erabilera. Egiaztatu da puntuazio horiek eraginkorrak direla hizketa eta isiltasun audio-bilbeak bereizteko. Atalase bat kalkulatuta soilik, oso emaitza lehiakorrak lortzen dira. Hala ere, VAD sistemak behar bezala funtziona dezan, audio-parametroak aurrez normalizatu behar dira, uneko audio-segmentuko (edo fitxategiko) batezbestekoak eta bariantzak erabiliz. Horrek *offline*koa bihurtzen du sistema, VAD analisia ezin baita hasi, datu guztiak prozesatu arte.

Arazo hori konpontzeko, atalasean oinarritutako teknikaren hedapen bat pentsatu dugu: normalizazio anitzeko puntuatzea edo MNS (*Multi-Normalisation Scoring*). Ideia horren funtsa da parametro akustikoak normalizatzea datu-multzo desberdinetatik erauzitako batezbestekoak eta bariantzak erabiliz; hala, sortzen diren behaketa-puntuazioen jokaera modelatu daiteke. Sailkatzaile gisa MLP bat erabiliz, atalasean oinarritutako teknikaz lortutako emaitzen antzekoak lortu ditugu. Hizketa garbirako, emaitzak oso lehiakorrak dira. Hizketa zaratatsurako, ordea, emaitzek funtzionamendu gorabeheratsua dute zarata motaren arabera. Entrenamenduan zenbait seinale zaratatsu erantsiz, arazoa leundu egiten da neurri batean. Izan ere, ER_1ari dagokionez, hobekuntzak lortzen dira SNR maila guztietan, baina nahiko altu jarraitzen dute mailarik zaratatsuenean (5 dB) lortzen diren errore-tasak. Seguruenik emaitza hobeak lortuko lirateke datu zaratatsu gehiago erabiliko balitz MLParen entrenamenduan.

Baliozkotze-esperimentu bat ere egin da, bi ITU-T VAD algoritmoren eta MNSan oinarritutako gure VADaren emaitzak alderatuz. Guk proposatutako VAD teknikaren funtzionamendua, hainbat SNRtako seinale zaratatsuekin (murmurio zarata eta zarata zuria) entrenaturik dagoenean, hobea da, oro har, ITU-T *G.720.1* eta *G.729b* sistema estandarren funtzionamendua baino, zeren sailkapen-errorea nabarmen txikiagoa baita isiltasunerako eta antzekoa da hizketa-segmentuetarako. Hori dela eta, gure teknika erabilgarria da, bai hizketa-errore baxua behar duten sistemetarako, bai isiltasun errore-tasa baxuak behar dituztenetarako. Gainera, gure VADak emaitzak behar bezala orokortzen dituela ematen du bitarteko SNRetarako eta ikusi gabeko zarata motetarako; eta horrek esan nahi du zarata-maila eta mota desberdinekiko sendoa dela sistema.

MNSan oinarritutako teknikaren abantailarik handienetariko bat da *online* funtzionatzen duela, erabakiak bilbez bilbe hartuz, inguruko bilbeak edo uneko segmentuko edo

fitxategiko bilbeak analizatu beharrik gabe. Gainera, VADaren oinarri gisa behaketa-egiantzak erabiltzea oso sinplea da, eta interesgarri bihurtzen du horrek. HMMak erabiltzen diren sistema batean (ASR sistema batean, adibidez), oso prozesamendu gehigarri gutxi behar du proposatutako VADak. Desabantailarik handiena izan daiteke nola funtzionatzen duen VADak, ikusi gabeko seinaleekin: badirudi emaitzak orokortzeko gai dela, baina errore-tasa zertxobait handiagotzen da SNR batzuetan. Ikerketa gehiago behar da zehazteko sistemak nola orokortu dezakeen modu egokian.

Etorkizuneko zenbait ikerketa-norabide izan litezke hizketa GMMtik lortutako behaketa-egiantzen analisia, emaitzen orokortzea aztertzea ikusi gabeko zaratak dituzten audio-seinaleak prozesatzean, MLPez gainera beste sailkatzaile desberdin batzuk ere probatzea —adibidez, neurona-sare errepikakorrak (RNN, *Recurrent Neural Networks*)— eta sistema ingurune errealeko baldintzatan testatzea.

CHAPTER 9

Online CMVNa

9.1 Sarrera

ASRen edo ASRan oinarritutako sistemen funtzionamendua andeatu egiten da entrenamenduko eta testeko seinaleen baldintza akustikoek ez badute bat egiten. Orain arte erabilitako teknikak bi kategoriatan sailkatu daitezke, oro har: eredu-egokitzapena eta parametro-normalizazioa. Ereduak egokitzeko tekniken funtsa da entrenatutako ereduak eraldatzea, testeko hizketaren baldintzekin bat egin dezaten; parametroak normalizatzeko teknikek, berriz, testeko parametroak aldatzen dituzte, entrenamenduko parametroen estatistikekin bat egin dezaten. Parametroak normalizatzeko teknikak, halaber, bi metodotan sailkatzen dira: metodo parametrikoak eta ez-parametrikoak. Tesi honetan, normalizazio parametrikoan jarriko dugu arreta, bereziki, batezbesteko- eta bariantza-normalizazio cepstrala (CMVN, *Cepstral Mean and Variance Normalisation*) teknikan.

Parametroak normalizatzeko, lehendabizi, batezbesteko-normalizazio cepstrala (CMN, *Cepstral Mean Normalisation*) proposatu zen [200]. CMNak esakuntza guztien lehen ordenako momentuaz bat egiten du, esakuntza bakoitzaren batezbestekoa zerora eraldatuz, denbora-batezbestekoa kenduta. Uste da kanalaren eragina konpentsatzen duela, zarata konboluzional gisa. 90. hamarkadaren amaieran, berriz, CMVN oso ezaguna bihurtu zen [201] [202] [203]. Bai batezbestekoak, bai bariantzak normalizatzen ditu, esakuntza guztiei zero batezbestekoa eta bateko bariantza izan dezaten eraldatuz. CMVNa asmatu zenetik, hainbat hobekuntza egin dira hartan oinarrituta, hala nola CMVN ereduaremuan erabiltzea [204], CMVNa SNR parametroekin konbinaturik erabiltzea [185], batezbestekoaren eta bariantzaren ondorengo estimazioak erabiltzea egiantz handieneko estimazioen ordez (ikuspegi Bayesiarra) [205], egiantz handieneko batezbesteko- eta bariantza-normalizazioa (ML-MVN, Maximum-Likelihood Mean and Variance Normalisation) erabiltzea esakuntza bateko hasierako batezbesteko- eta bariantza-bektoreak estimatzeko (CMVN geroago aplikatzen da, bilbe kopuru finko batetik aurrera) [206] eta abar. Teknika berri bakoitzak bere abantailak eta desabantailak ditu, baina gaur egun oso ohikoa da CMVN aplikatzea hizketa-ezagutze sendoaren eremuan [207].

Hurrengo atalean (9.2. atala) CMVNren oinarriei buruz jardungo dugu. Zaratak eta kanal-distortsioak hizketa garbian duten eragina aztertuko da. 9.3. atalean, CMVNaren *online*ko hiru inplementazio deskribatzen dira, eguneratze kausal errekurtsiboa darabiltenak. Inplementazio bakoitzaren emaitza esperimentalak ere ageri dira. 9.4. atalean, *online* normalizazio-metodo berri bat aurkezten da (8.4. atalean azaldutako MNS teknikarekin lotura zuzena duena): MNSan oinarritutako CMVNa. Azkenik, zenbait konklusio azaltzen dira 9.5. atalean.

9.2 CMVNren oinarriak

CMVN aplikatzearen esanahia ulertzeko, demagun sarreran $x[n]$ seinalea dugula eta kanalaren pultsu-erantzuna $h[n]$ dela. Azken seinalea bien konboluzio lineala da (ikus (9.1) ekuazioa).

$$y[n] = x[n] * h[n] \tag{9.1}$$

Fourierren Transformatua hartuta, haren konboluzio- eta biderkaketa-baliokidetasuna dela eta, (9.2) ekuazioa lortzen dugu.

$$Y[f] = X[f] \cdot H[f] \tag{9.2}$$

Cesptrumak kalkulatzeko, espektroaren logaritmoa kalkulatzen dugu (ikus (9.3) ekuazioa).

$$\log Y[f] = \log\left(X[f] \cdot H[f]\right) = \log X[f] + \log H[f] \tag{9.3}$$

Eta orain denboraren eremura bueltatzen gara berriro (edo, hobe esanda, q *quefrency*-ren eremura) alderantzizko FTaren bidez. Hala, cepstrum koefizienteak lortzen dira (ikus (9.4) ekuazioa).

$$Y[q] = X[q] + H[q] \tag{9.4}$$

Ikus daitekeenez, eremu cepstralean, erantsiz adierazten dira distortsio konboluzionalak. Jo dezagun hizketa-seinaleak jasaten dituen distortsioak geldikorrak direla (nahiko baieztapen errealista da, zeren ahots-bidea eta kanal-erantzuna ia-ia ez baitira aldatzen denbora-segmentu oso txikietan zehar) eta hizketa-seinalea seinale zatika geldikor gisa ikus dezakegula (edo epe laburreko seinale geldikor gisa). Orduan, $i.$ bilbe (geldikor)

guztietarako, (9.5) ekuazioa betetzen da.

$$Y_i[q] = X_i[q] + H[q] \tag{9.5}$$

Bilbe guztietan zehar batezbestekoa eginez, (9.6) ekuazioa lortzen dugu.

$$\frac{1}{N}\sum_i Y_i[q] = \frac{1}{N}\sum_i X_i[q] + H[q] \tag{9.6}$$

Orain, i. bilbeko balio cepstralaren (Y_i) eta batezbestekoaren aldeari D_i baderitzogu, (9.7) ekuazioan bezala idatz daiteke.

$$\begin{aligned} D_i[q] = Y_i[q] - \frac{1}{N}\sum_j Y_j[q] = X_i[q] + H[q] - \left[\frac{1}{N}\sum_j X_j[q] + H[q]\right] = \\ = X_i[q] - \frac{1}{N}\textstyle\sum_j X_j[q] \end{aligned} \tag{9.7}$$

(9.7) ekuazioari begiratuta, ondoriozta daiteke ezen, cepstrumetatik batezbesteko balioak kenduta, kanalaren distortsioa ezabatzen dugula.

Testuinguru zaratatsuetarako, zarata gehigarria dela kontsideratuta, (9.1) ekuazioa (9.8) ekuazioa bihurtzen da, eta (9.2) ekuazioa, berriz, (9.9) ekuazioa.

$$y[n] = x[n] * h[n] + w[n] \tag{9.8}$$

$$Y[f] = X[f] \cdot H[f] + W[f] \tag{9.9}$$

Eta logaritmoa kalkulatuz, (9.10) ekuazioa lortzen dugu.

$$\log Y[f] = \log\left[X[f]\left(H[f] + \frac{W[f]}{X[f]}\right)\right] = \log X[f] + \log\left(H[f] + \frac{W[f]}{X[f]}\right) \tag{9.10}$$

Orain, $\frac{W[f]}{X[f]}$ terminoa ageri da, zeina mespretxagarria izan baitaiteke zarata baxuko baldintzetan, baina, era berean, eragin handikoa, SNR txikiko baldintzetan.

Laburbilduz, CMN gai da distortsio konboluzionalak konpentsatzeko. Are gehiago, frogatu da batezbesteko cepstralak karakterizatzen duela, kanal-transferentziaren funtzioa ez ezik, hizlari desberdinen ahots-bideen maiztasun-erantzunaren batezbestekoa ere. Ondorioz, hizlarien epe luzeko batezbestekoa kenduz, hizlari-normalizatzaile gisa jokatu dezake CMNak [133]. Bestalde, CVN ez dago loturik distortsio mota jakin bat konpentsatzearekin; hala ere, sendotasuna ematen du kanal akustikoen desberdintasunen, hizlarien aldakortasunen eta zarata gehigarriaren kontra [201]. CMVNak bi tekniken abantailak batzen ditu.

Ikuspegi ohikoenean, parametroen batezbesteko- eta bariantza-bektoreak estimatzeko, esakuntza bat [183] edo esakuntza batetik leihokatutako atal bat (eskuarki, denbora errealeko sistemetarako) [203] erabiltzen da. Orduan, parametro-bektoreak desplazatu eta eskalatu egiten dira estimatutako batezbestekoez eta bariantzez, helburua izanik parametro normalizatuek zero batezbestekoa eta bateko bariantza izatea. N bektore cepstraletarako: $y = \{y_1, y_2, ..., y_N\}$, μ batezbestekoen eta σ^2 bariantzen bektoreak (9.11) ekuazioan eta (9.12) ekuazioan adierazi bezala kalkulatzen dira, hurrenez hurren.

$$\mu_N(i) = \frac{1}{N}\sum_{n=1}^{N} y_n(i) \tag{9.11}$$

$$\sigma_N^2(i) = \frac{1}{N}\sum_{n=1}^{N} (y_n(i) - \mu_N(i))^2 \tag{9.12}$$

non i: bektorearen i. osagaia. Cepstrum parametroak, orduan, normalizatu egiten dira, kalkulatutako batezbesteko- eta bariantza-bektoreak erabiliz, (9.13) ekuazioan adierazten den moduan.

$$\hat{y}_n(i) = \frac{y_n(i) - \mu_N(i)}{\sigma_N(i)} \tag{9.13}$$

9.3 *Online*ko inplementaziorako CMVNren azterketa

CMVNarekin lortzen diren ezagutze-emaitzarik onenak, eskuarki, MFCCen batezbestekoen eta bariantzen balioak aurrez kalkulatuz lortzen dira uneko esakuntza osoa erabiliz, eta konstante uzten dira normalizazio-prozesu osoan zehar [208]. Ordenagailu Bidezko

Ebakera Lanketa (OBEL) atazan, audio-laginak paketatuta bidal daitezke wav fitxategi batean; hala, CMVN erraz aplikatu daiteke, hasieran fitxategi osoa prozesatuz eta gero fitxategiko parametro cepstralak normalizatuz. Hala eta guztiz ere, Hitzez Hitzeko Esaldi Egiaztapenean (HHEE), audio-laginak iritsi ahala aplikatu behar da normalizazioa, eta, beraz, batezbestekoak eta bariantzak ezin dira kalkulatu audio-segmentu osoa erabiliz. Hasierako estimazioa behar da, bai batezbestekoentzat, bai bariantzentzat, eta, orobat, ondorengo eguneraketa edo egokitzapena.

Metodo desberdinak ageri dira literaturan, cepstrum parametroak *online* normalizatzeko:

- **Iraganeko datuak erabiltzea**: Metodologia honetan, parametroen batezbestekoak eta bariantzak ez dira estimatzen uneko esakuntza erabiliz, baizik eta beste datuiturri batzuk; hala nola uneko saio edo saio-multzo bereko seinaleak, edo baita prozesatutako azken seinalea ere. Gero, balio konstante horiekin normalizatzen dira seinaleak. Metodo horrekin oso emaitza onak lortzen dira, [209]en ageri denez, baina badu desabantaila bat: iraganeko datuak behar dira.
- **Ikuspegi segmentala**: Lehendabizikoz [203]en proposatu zen, non, MFCC bektoreak normalizatzeko, luzera finituko normalizazio-leiho lerrakor bat erabiltzen baita, uneko bilbea erdian kokaturik duena, eta CMVNa aplikatzen baitzaio. Hitz isolatuko ezagutze-esperimentuetan, 1 s inguruko leihoekin lortzen dira emaitzarik onenak; horrek esan nahi du 0.5 s-ko atzerapena eransten zaiola, eta hori luzeegia izan daiteke HHEE sistema batean aplikatzeko.
- **Ikuspegi errekurtsiboa**: Batezbesteko- eta bariantza-bektoreak uneko esakuntzako lehen D bilbeak erabiliz hasieratzen dira, eta, gero, bilbe gehiago iritsi ahala, errekurtsiboki eguneratzen dira [210]. Hortaz, sistema ez-kausala da, estimazioa eguneratzeko erabilitako bilbea D bilbe aurrerago baitago normalizatu beharreko bilbea baino (horretxegatik esaten zaio D aurrera begirako parametroa), (9.14) ekuazioan eta (9.15) ekuazioan ageri denez. Bistan denez, D da normalizazio-teknika horrek eragiten duen atzerapena. Kontuan izan $x_n(i)$ zera dela, i. parametro edo bektore-osagaia n bilbean; β, berriz, ahazte-faktorea.

$$\mu_n(i) = \beta \cdot \mu_{n-1}(i) + (1-\beta) \cdot x_{n+D}(i) \tag{9.14}$$

$$\sigma_n^2(i) = \beta \cdot \sigma_{n-1}^2(i) + (1-\beta) \cdot (x_{n+D}(i) - \mu_n(i))^2 \tag{9.15}$$

Atzerapen bererako, ikuspegi segmentalaz baino emaitza hobeak lortzen dira ikuspegi errekurtsiboaz.

Ikuspegi errekurtsiboa hobetu egin daiteke, batezbestekoen eta bariantzen hasierako balioak hobeto estimatuz. Adibidez, hasierako balioak estimatzeko iraganeko datuak erabiliz eta gero errekurtsiboki eguneratuz, *offline*ko emaitzen oso antzekoak lor daitezke. Iraganeko daturik erabili ezin bada, lehen isiltasun-hizketa trantsizioa erabil daiteke, esakuntzaren hasierako segmentuak erabiltzeak baino emaitza hobeak lortzen baitira [209]. Dena dela, horretarako, VAD bat behar da.

9.3.1 *Online* CMVNaren zenbait inplementazio

Atal honetan, hiru *online* normalizazio-teknikaren eragina ikusiko dugu. Lehenengoan, hasierako batezbestekoak eta bariantzak lehen D bilbeak erabiliz kalkulatzen dira, eta, gero, datozen bilbeak erabiliz eguneratzen dira. Bigarrenean, hasierako balioak aurrez estimatzen dira, entrenamenduko datu-basetik. Hirugarrenean, teknika hibrido bat erabiltzen da, non batezbestekoen hasierako balioak lehen D bilbeak erabiliz estimatzen diren eta gero bilbez bilbe eguneratzen diren eta bariantzak, berriz, entrenamenduko datu-basetik estimatzen diren eta konstante eusten zaien.

a) Hasierako aurrera begirakoa eta eguneratze errekurtsiboa

Teknika honetan, batezbestekoen eta bariantzen balioak lehen D bilbeak erabiliz kalkulatzen dira. Estimatutako batezbesteko eta bariantza horiek erabiliz, D bilbeak normalizatzen dira, eta, hortik aurrera, uneko bilbea erabiliz eguneratzen dira batezbestekoak eta bariantzak, (9.16) ekuazioan eta (9.17) ekuazioan adierazten den bezala.

$$\mu_n(i) = \frac{(n-1)\mu_{n-1}(i) + x_n(i)}{n} \tag{9.16}$$

$$\sigma_n^2(i) = \frac{(n-1)\sigma_{n-1}^2(i) + (x_n(i) - \mu_{n-1}(i))(x_n(i) - \mu_n(i))}{n} \tag{9.17}$$

9.1. irudian, teknika hori erabiltzearen ondorioen adibide bat ageri da: goiko irudian, bigarren parametroaren ($c1$) balioak ageri dira denboran zehar (10 ms oro jasotako bilbeak), batezbestekoaren *offline* balioarekin (konstantea) eta errekurtsiboki eguneratzen den batezbestekoaren *online* balioarekin batera. Erdian, *offline* bariantza-kurbak eta *online* kurbak. Behean, $c1$ kurba normalizatuak *offline* kasuan eta *online* kasuan.

Irudiak erakusten du oso garrantzitsua dela hasieran batezbestekoak eta bariantzak ondo estimatzea, parametroak behar bezala normalizatzeko. Bariantzaren hasierako balioa isiltasun-bilbeak erabiliz bakarrik kalkulatuz gero, oso balio baxua lortuko da, eta horrek eragin handia izango du koefiziente-balioak normalizatzean. Batezbestekoaren kasuan, ez dirudi halako eragin handia duenik.

Esperimentu fonetikoaren emaitzak 9.3.2. atalean aurkezten dira.

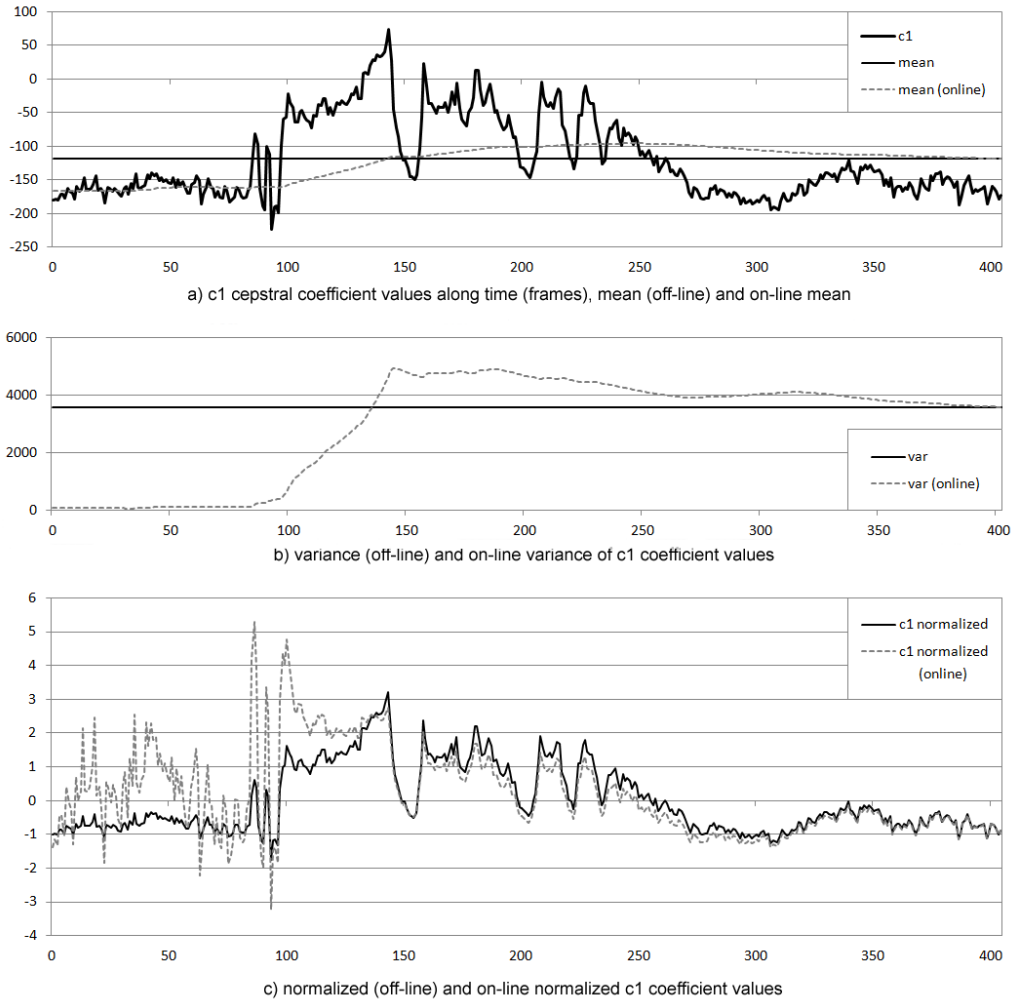


Figure 9.1: Hasierako aurrera begirakoa eta eguneratze errekurtsiboa: a) $c1$ koefizientea (kurba lodi beltza), *offline* batezbestekoa (beltza) eta *online* batezbestekoa (etena) 250 ms-ko aurrera begirakoarekin; b) $c1$ *offline* bariantza (beltza) eta *online* bariantza (etena); c) $c1$-en balio normalizatuak *offline* (beltza) eta *online* (etena) moduetarako.

b) Iraganeko datuak erabiltzea

Iraganeko datuak erabiltzeak, batezbestekoen eta bariantzen hasierako balioak estimatzeko, oso eraginkorra dirudi uneko audio seinalearen ezaugarriak aurrekoen antzekoak baldin badira. Balio horiek lortzeko ohiko modu bat da uneko saioa edo saio multzoa erabiltzea. Uneko saioa (erabiltzaile berari dagozkion seinaleak) erabilita emaitza hobeak lortzen badira ere, batez ere distantzia ertainetarako, hor dirau hasieratzearen arazoak. Saio multzo bat erabilita emaitza pittin bat okerragoak lortzen dira, baina datu gehiago daude egon liteke erabilgarri [209].

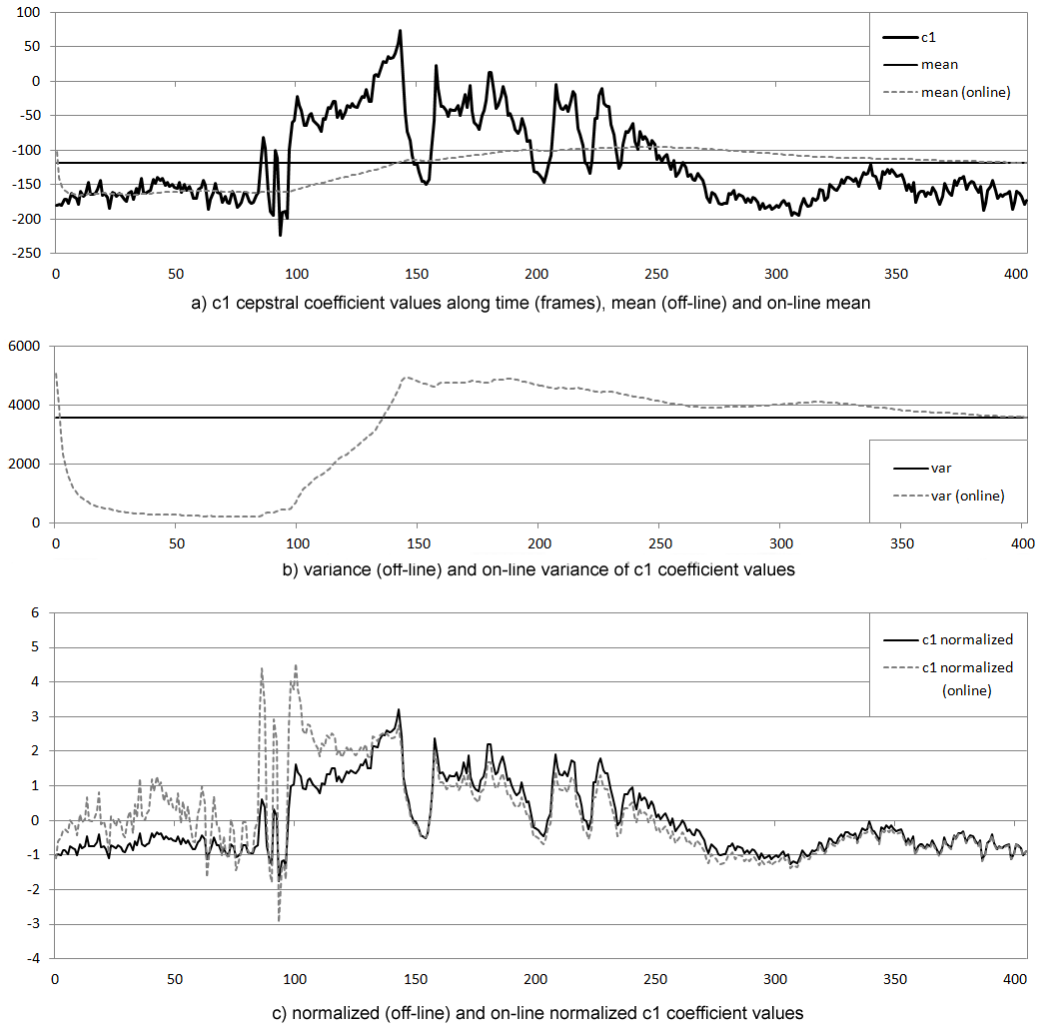


Figure 9.2: Iraganeko datuak hasierako estimazio gisa, eta eguneratze errekurtsiboa: a) $c1$ koefizientea (kurba lodi beltza), *offline* batezbestekoa (beltza) eta *online* batezbestekoa (etena); b) $c1$ *offline* bariantza (beltza) eta *online* bariantza (etena); c) $c1$-en balio normalizatuak *offline* (beltza) eta *online* (etena) moduetarako.

Atal honetan, batezbestekoen eta bariantzen balio orokorrak estimatuko ditugu, *Basque Speecon-like* datu-basea erabiliz. Bi eszenatoki kontsideratu dira:

- **Konstante-modua**: aurrez kalkulatutako balioei konstante eusten zaie. Emaitzak oso intuitiboak dira: koefiziente normalizatuen kurba lerratu eta anplifikatu egingo da, proportzionalki, entrenamenduko seinaleen batezbestekoen eta bariantzen eta uneko audio seinaleenen arteko aldeekiko.

- **Eguneratze-modua**: aurrez kalkulatutako balioak errekurtsiboki eguneratzen dira unean analizatzen ari den bilbeaz. Beharrezkoa da egiaztatzea ea koefiziente normalizatuak egonkorrak diren ala ez.

9.2. irudiak eguneratze-moduaren adibide bat ilustratzen du grafikoki. Aurrez kalkulatutako bai batezbestekoaren, bai bariantzaren hasierako balioen eraginak oso denboratarte laburra irauten du: batezbesteko-kurbak berehala bat egiten du *online* kurbarekin, bilbe gutxitara; bariantza-kurbak zenbait bilbe gehiago behar ditu, baina berehala erortzen da zero ingurura, eta horrek esan nahi du koefiziente-balioak gehiegi ari direla anplifikatzen. Horrek aurreko inplementaziora garamatza, eta, hortaz, antzeko emaitzak espero dira.

Iraganeko datuak *konstante*-moduan nahiz *eguneratze*-moduan erabiliz egindako esperimentu fonetikoen emaitzak 9.3.2. atalean aurkezten dira.

c) Teknika hibridoa

Aurreko bi inplementazioetan, agerian geratu da azken emaitzetan eragin kaltegarria duela batezbestekoen eta bariantzen hasierako balioak txarto estimatzeak. Are gehiago, bariantza behar bezala kalkulatzea erabakigarriagoa da, zeren koefiziente normalizatuak desbideratze estandarraz (bariantzaren erro karratuaz) zatituz kalkulatzen baitira.

Ikuspegi berri honetan, aurreko bi inplementazioen ezaugarriak erabiltzen dira, hortik haren izena: teknika hibridoa. Alde batetik, aurrera begirako teknika baliatzen du batezbestekoen hasierako balioak estimatzeko. Aurrera begirako D bilbeen ondoren, batezbestekoak errekurtsiboki eguneratzen dira. Beste aldetik, iraganeko bariantza-datuak erabiltzen dira konstante-moduan, eguneratu gabe. 9.3. irudiak grafikoki adierazten du hori guztia.

Teknika hibridoan, konpondu egiten da bariantzen hasierako balio baxuen arazoa, eta horrek hasierako distortsioak saihesten laguntzen du. Era berean, batezbesteko-balioak lehen D bilbeetatik estimatzen dira eta eguneratuz joaten dira handik aurrera. Horrekin, cepstrumen normalizazio nahiko uniformea bermatzen da, hasieran izan ezik, non balitekeen koefiziente normalizatuak lerraturik agertzea batezbestekoen arteko aldea dela eta. Dena dela, ez dirudi horrek eragin handia duenik isiltasuna ezagutzeko unean; are gehiago, lehen D bilbeetan hizketa-bilbeak baldin badaude, batezbestekoen estimazio hobea lortuko da. Teknika honen alderik txarrena da uneko audio-seinalearen bariantza ez dela oso desberdina izan behar aurrez kalkulatutakoarekin alderatuta.

Teknika hibridoaz egindako esperimentu fonetikoaren emaitzak hurrengo atalean ageri dira.

9.3.2 Emaitza esperimentalak

Hasierako batezbestekoak eta bariantzak estimatzeko *online* metodoarekin sistemari gehitzen zaion distortsioa ebaluatzeko, errepikatu egin da 5.3. atalean egindako test fonetikoa, baina aurreko atalean deskribatutako *online* inplementazio desberdinak erabiliz, *offline*koa erabili beharrean. Testerako erabilitako HMMak *offline* testean emaitzarik

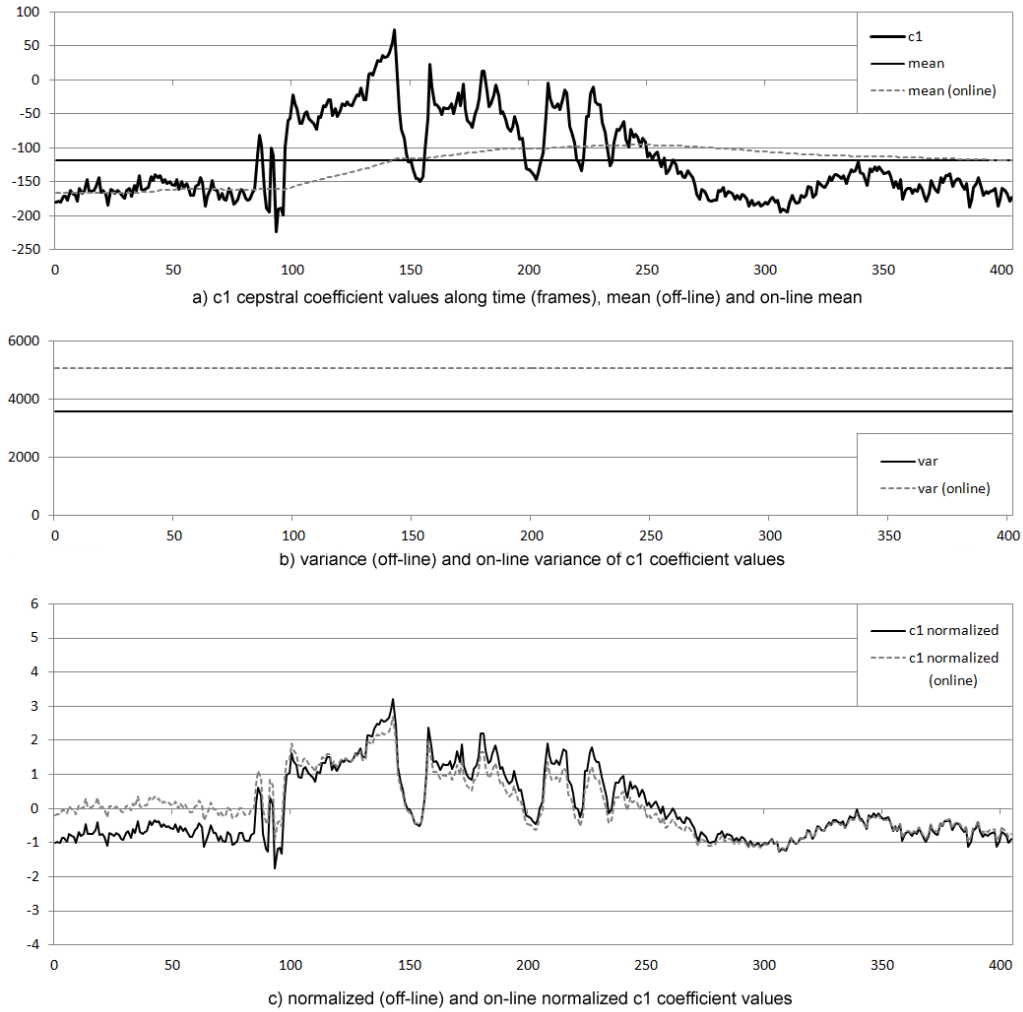


Figure 9.3: Teknika hibridoa: a) $c1$ koefizientea (kurba lodi beltza), *offline* batezbestekoa (beltza) eta *online* batezbestekoa (etena); b) $c1$ *offline* bariantza (beltza) eta *online* bariantza (etena); c) $c1$-en balio normalizatuak *offline* (beltza) eta *online* (etena) moduetarako.

onenak izandakoak dira. Hurbileko azpicorpuseko audio-fitxategiak (entzungailu mikrofonodun batez grabatuak) eta mahai gaineko azpicorpuseko fitxategiak (1 m-ko distantzian kokatutako mikrofono batez grabatuak) testatu dira, ikuspegi zabalago bat osatzearren. 9.1. taulan, test horietan lortutako emaitzak ageri dira, *online* inplementazio bakoitzerako eta kanal desberdin bakoitzerako, fonemen errore-tasa (PER, *Phone Error Rate* %) eta Zehaztasuna (%) baliatuta. Taulan *offline* emaitzak ere jarri dira, irakurlearentzat emaitzak interpretatzea errazagoa izan litekeelakoan.

9.1. taularen emaitzek erakusten dute ezen, espero bezala, PERak eta Zehaztasunak,

Table 9.1: Hiru *online* inplementazioen emaitzak: PERak eta Zehaztasunak, hurbileko eta 1 m-ko distantziako seinaleentzat, *offline* balioekin alderatuta.

	PER (%)				
	Offline	*Online*			
	Uneko datuak	Aurr. begirakoa + egun.	Irag.datuak	Irag. datuak + egun.	Hibridoa
Hurbilekoa	12.45	18.39	14.20	18.29	15.30
Mahai gainekoa	21.87	29.92	29.98	29.62	33.38

	Zehaztasuna (%)				
	Offline	*Online*			
	Uneko datuak	Aurr. begirakoa + egun.	Irag.datuak	Irag. datuak + egun.	Hibridoa
Hurbilekoa	74.24	38.91	59.00	42.09	54.05
Mahai gainekoa	60.12	22.13	37.04	26.72	43.79

online kasu guztietan, *offline* kasuetakoak baino okerragoak direla. Iraganeko datuak erabilita aurrez kalkulatutako batezbesteko eta bariantza horiei seinale osorako konstante eutsiz, PER pittin bat okerragoak lortzen dira hurbileko fitxategiak testatzean, baina askoz okerragoak mahai gaineko fitxategiak testatzean, kanal-desberdintasuna dela eta. Zehaztasuna ere, zeinak txertaketak, ordezkapenak eta ezabaketak kontuan hartzen baititu, nabarmen okerragoa da bi kasuetan, batez ere mahai gaineko kanaleko fitxategietarako, narriadura handia lortzen da eta. Dena dela, hain emaitza eskasak lortuta ere, teknika horrekin lortzen dira emaitzarik onenak hurbileko fitxategietarako.

Eguneratzea duten teknikekin (aurrera begirakoa + eguneratzea eta iraganeko datuak + eguneratzea), oso antzeko emaitzak lortzen dira. Zehaztasunari dagokionez, oso emaitza baxuak lortzen dira bi tekniketarako (nahiz eta iraganeko datuak + eguneratzea teknikaz emaitza pittin bat hobeak lortzen diren). Okerragotze hori hasieran gertatzen da batez ere, bariantzak kalkulatzean isiltasun-bilbeak soilik kontsideratzen baitira. Horren ondorio gisa, fonema-txertaketak gertatzen dira isiltasunaren ordez, zeren hasierako bilbe horiek indartu egiten baitira isiltasun-balioen bariantza baxuaren eraginez.

Teknika hibridoaz lortzen dira emaitzarik onenak uneko datuak erabiltzen dituztenen tekniken artean, mahai gaineko fitxategien PERetan izan ezik. Bigarren PER-rik onena lortu du hurbileko kanaleko fitxategiak testatzean, baina okerrenak, mahai gaineko fitxategietan. Zehaztasunari dagokionez, bigarren emaitzarik onena lortu da hurbileko fitxategietarako ere, eta onena, mahai gaineko fitxategietarako; aski eskasak, halere.

Laburbilduz, iraganeko datuak erabiltzea arazoa konpontzeko konponbide sinplea da, baina emaitzak nabarmen okerragotzen dira kanal desberdin batean zehar grabatutako seinaleak testatuz gero, ez baita inolako egokitzapenik egiten uneko seinalearen ezaugarrietara. Bestalde, batezbestekoak eta bariantzak bilbez bilbe eguneratuta, ez dira oso emaitza onak lortzen, batez ere bai isiltasun-bilbeak, bai hizketa-bilbeak erabili behar baitira batezbestekoak eta bariantzak behar bezala estimatzeko. Teknika hibrido bat ere testatu da, non, batetik, batezbestekoen hasierako balioak estimatzeko aurrera begirakoa erabiltzen den, eta, aurrera begirako D bilbeen ondoren, errekurtsiboki eguneratzen diren, bestetik, iraganeko bariantza-datuak konstante-moduan erabiltzen dira, eguneratzerik gabe. Teknika horrek aurreko bi tekniken alde onak aprobetxatu beharko lituzke, baina, hala ere, emaitzak ez dira onak.

9.4 MNSan oinarritutako CMVNa

9.4.1 Sarrera

8.4. atalean, VAD teknika berri bat aurkeztu dugu, tesi honetan proposatutako MNS metodoan oinarritua. Metodo horren funtsa da hainbat behaketa-egiantz sortzea uneko MFCCak normalizatuta, baldintza desberdinetan grabatutako hainbat hizketa datu-multzotatik kalkulatutako batezbestekoak eta bariantzak erabiliz. Hala, behaketa-egiantzen multzo (edo bektore) bat lortzen da, patroi bat edo beste bat duena, prozesatzen ari den bilbearen izaeraren ("hizketa" izan edo "isiltasuna" izan) eta SNRaren arabera.

VADrako, geruza anitzeko pertzeptroia (MLP, *Multi-Layer Perceptron*) erabili da, MNSaz lortutako puntuazio-bektoreak bi klasetan sailkatzeko: *hizketa* eta *isiltasuna*. Hala ere, helburua, orain, bestelakoa da: aurkitu behar da zer datu-multzori dagokion bilbe bakoitza (eta horrek ez luke atzerapenik eragingo). MLPak informazio hori emango baligu, datu-multzo horretatik kalkulatutako batezbesteko eta bariantza orokorrak erabili ahal izango genituzke seinale berriaren MFCCak normalizatzeko. Hala, jotzen badugu N datu-base ditugula eskura —bakoitzak bere MFCC μ_n batezbesteko-bektorea eta σ_n^2 bariantza bektorea duela—, MLP berriaren emaitza izango da bilbe bakoitzak zer probabilitate duen datu-multzo bakoitzekoa izatekoa. MNSan oinarritutako CMVNaren ideia nagusia da sarrerako bilbe bakoitzaren batezbestekoak eta bariantzak estimatzea (bektorearen i. osagai bakoitzeko, $\hat{\mu}(i)$ eta $\hat{\sigma}^2(i)$, hurrenez hurren), datu-multzo bakoitzeko batezbestekoen eta bariantzen batura haztatu gisa, probabilitateak w_n haztapen gisa erabiliz. Hori (9.18) ekuazioan eta (9.19) ekuazioan ageri da.

$$\hat{\mu}(i) = \sum_{n=1}^{N} w_n \cdot \mu_n(i) \tag{9.18}$$

$$\hat{\sigma}^2(i) = \sum_{n=1}^{N} w_n \cdot \sigma_n^2(i) \tag{9.19}$$

MLP berriak sarrera eta irteera kopuru berbera izango du. MLParen sarrera N elementuko puntuazio bektore bat izango da; irteera, berriz, erabiltzen ari garen N datu-multzo bakoitzekoa izateko probabilitateak izango dira.

9.4.2 MNSan oinarritutako CMVN esperimentuak

Proposatu dugun MNSan oinarritutako CMVN metodoaren baliozkotasuna egiaztatzeko, ebaluatu behar dugu ea estimatutako MFCC batezbestekoak eta bariantzak egokiak diren. Lehendabizi, hasierako ezagutze fonetikoko esperimentu bat egin dugu, batezbestekoak eta bariantzak sarreran azaltzen den bezala estimatuz. Datu-multzo desberdinak kontsideratu dira MNSan MLPa entrenatzeko, MLP eredu orokor bat sortzeko asmoz.

Hasierako esperimentuaren ondoren, emaitzen analisia egin da, ikusteko zein diren isiltasun-bilbeak soilik erabiliz estimatutako batezbestekoen eta bariantzen balioen eta bilbe guztiak erabiliz estimatutakoen artean. Horrek beste esperimentu bat egitera eraman gaitu: *Basque Speecon-like* datu-baseko *hurbileko* kanala erabiliz entrenatutako isiltasun HMMaren erdiko egoeratik lortutako behaketa-egiantzak ez ezik, hizketa-datu guztiekin entrenatutako GMM berri batetik lortutakoak ere hartu dira kontuan. Hala, MLP berri bat entrenatu da, eta berriro egin da hasierako esperimentua.

Azkenik zenbait ondorio ikusiko ditugu.

a) Ezagutze fonetikoko hasierako esperimentua

Hasierako esperimenturako, MLP bat entrenatu da, MNS metodoa 12 datu-multzotan ezarriz lortutako behaketa-egiantzekin: *Basque Speecon-like* datu-baseko *hurbileko* eta *mahai gaineko* kanalak [99], *Spanish Speecon* datu-baseko azpi-multzoko C_1 eta C_3 kanalak, ahots-aktibitatea eta ahostuntzea detektatzeko ECESS ebaluazio-kanpainan erabilia [180] eta *Noisy TIMIT* hizketa datu-baseko 8 datu-multzo [197], hain zuzen ere murmurio-zarataren kanaleko 4 azpimultzo eta zarata zuriaren kanaleko beste 4: 50, 35, 20 eta 5 dB balioko SNR balioei dagozkienak. Datu-multzo guztiak dira datu-baseetako *Train* blokeetakoak; *Test* atala esperimentuetarako utzi da.

Ezagutze fonetikorako erabilitako HMMak *R+M25* prozesuari jarraituz sortutakoak dira, hiztegi alternatibarik gabea eta 32 gaussiar erabiliz. Kontuan izan % 12.45eko PERa zela HMM horiei *offline* CMVNa aplikatuz lortutako emaitzarik onena (ikus 5.3. atala).

Esperimentuen emaitzak 9.1. taulan ageri dira. 9.1. taulan aurkezten diren balioekin alderatuta, esperimentu honetan *hurbileko* azpi-multzorako lortutako PERa hirugarren onena da; *mahai gaineko* azpi-multzorako lortutako PERa, aldiz, baxuena da. Zehaztasunei dagokienez, esperimentu honetan lortutako emaitzak dira baliorik onenak.

Table 9.2: MNSan oinarritutako CMVNaren inplementazioaren (*online*) emaitzak: PERak eta Zehaztasunak, *hurbileko* eta *mahai gaineko* seinaleekin, *offline* balioekin alderatuta.

	PER (%)	
	Offline	*Online*
	Uneko datuak	MNSan oinarritua
Hurbilekoa	12.45	15.51
Mahai gainekoa	21.87	29.29

	Zehaztasuna (%)	
	Offline	*Online*
	Uneko datuak	MNSan oinarritua
Hurbilekoa	74.24	65.50
Mahai gainekoa	60.12	45.45

b) Datuen analisia

Estimatutako batezbesteko eta bariantzen eta benetakoen arteko antzekotasuna aztertzeko, beste esperimentu bat egin dugu: MLP berri bat entrenatu dugu 8 datu-multzotatik lortutako behaketa-egiantzen bektoreak erabiliz: *Noisy TIMIT* datu-baseko murmurio-zaratako eta zarata zuriko datu-multzoetako *Train* blokeetako 50, 35, 20 eta 5 dB-ko kanalak. MLP horrekin, kanal guztiak (Test blokekoak) testatu dira: 50 dB-tik 5 dB-ra, 5 dB-ko tartetan. Hala, sarrerako bektore bakoitzak 8 datu-multzoetariko bakoitzekoa izateko duen probabilitatea kalkulatu dugu, eta, gero, (9.18) ekuazioa eta (9.19) ekuazioa aplikatu ditugu MFCCen batezbestekoak eta bariantzak estimatzeko.

Bi test egin dira: alde batetik, isiltasun-bilbeak soilik hartu dira kontuan MFCCen batezbestekoak eta bariantzak estimatzeko. Bestetik, bilbe guztiak erabili dira MFCCen batezbestekoak eta bariantzak estimatzeko. Bereizketa hori egitearen zergatia da ezen, behaketa-egiantzak isiltasun GMM batez kalkulaturik daudenez, zentzuzkoa dela pentsatzea MLPak hobeto sailkatuko dituela isiltasun-bilbeak.

9.4. irudian, ageri dira 0. MFCCa (goiko diagramak) eta 1. MFCCa (beheko diagramak), murmurio datu-multzoko zarata-maila guztietarako. Irudietako kurba berdeez, testatutako fitxategien batezbesteko errealen balioak adierazten dira; kurba laranjez, estimatutako batezbestekoen batez besteko balioak adierazten dira, testatutako datu-multzoetako bilbe guztiak kontuan hartuta; kurba gorriez, berriz, estimatutako batezbestekoen batez besteko balioak, baina testatutako datu-multzoetako isiltasun-bilbeak soilik kontuan hartuta. Bistan da isiltasun-bilbeak soilik erabiliz lortzen diren kurbak balio errealen kurben antzekoagoak direla. Horren azalpena izan liteke ezen

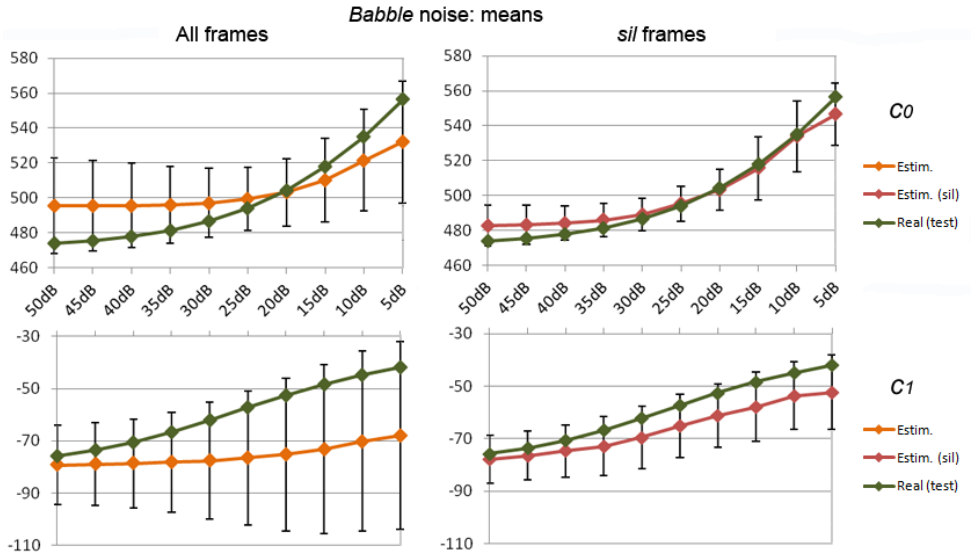


Figure 9.4: 0. (goian) eta 1. (behean) MFCCen batezbestekoen balio errealak vs. balio estimatuak, eta desbideratze estandarrak (ezkerrean: bilbe guztiak erabiliz; eskuinean: isiltasun-bilbeak erabiliz), murmurio datu-multzoko zarata-maila guztietarako.

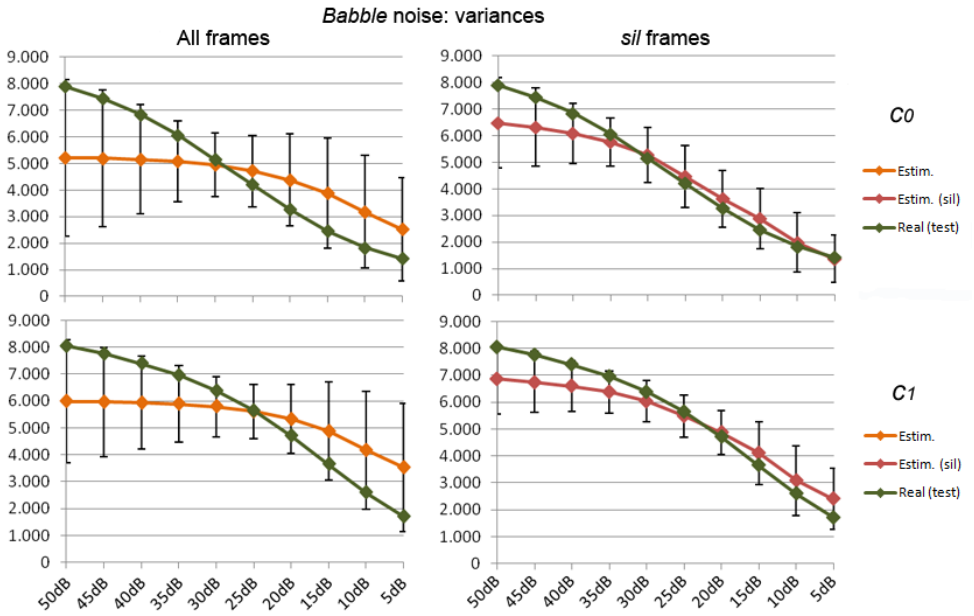


Figure 9.5: 0. (goian) eta 1. (behean) MFCCen bariantzen balio errealak vs. balio estimatuak, eta desbideratze estandarrak (ezkerrean: bilbe guztiak erabiliz; eskuinean: isiltasun-bilbeak erabiliz), murmurio datu-multzoko zarata-maila guztietarako.

MNS puntuazio-bektoreek gorabehera handiagoak dituztela hizketa-bilbeetan isiltasun-bilbeetan baino (ikus 8.8. irudia, 8.4. atalean), eta, horren ondorioz, zailagoa da hizketa-bilbeei dagozkien bektoreak modelatzea.

9.5. irudian, 0. eta 1. MFCCen bariantzak, errealak eta estimatuak, ageri dira, murmurio zaratarako. Emaitzek konklusio berera garamatzate: bariantzen balio errealen kurbak (berdez) eta balio estimatuenak alderatuz gero (bilbe guztiak erabiliz (laranjaz) eta isiltasun-bilbeak soilik erabiliz (gorriz)), isiltasun-bilbeak soilik erabiliz lortzen diren kurbek antz handiagoa dute balio errealen kurbekin.

Ondorio bera atera dugu zarata zuriaren kanaleko seinaleetarako.

c) Ezagutze fonetikoko esperimentua, *hizketa* GMMa erantsiz

Hizketa-bilbeak hobeto modela daitezkeen ikusteko, hizketa GMM batek sortutako behaketa-egiantzak ere behar dira. Gure behaketa-egiantzen bektoreak isiltasun GMMa erabiliz sortuak dira, *Basque Speecon-like* datu-baseko hurbileko kanalaz entrenatua. Halaber, hizketa GMM bat entrenatu da, *Basque Speecon-like* datu-baseko hurbileko kanaleko *Train* blokeko hizketa-bilbe guztiak erabiliz.

MLP berri bat entrenatu da, hasierako esperimentuko datu-multzo berberak erabiliz. MLP hori entrenatzeko bektoreek, orain, isiltasun GMMak emandako 12 behaketa-egiantz eta hizketa GMMak emandako beste 12 dituzte. Horrenbestez, MLP berriak 24 sarrera, 12 irteera (erabilitako datu-multzo bakoitzeko bat) eta 18 neuronako ezkutuko geruza bat ditu. MLP berri horrekin, errepikatu egin da hasierako esperimentua. Emaitzak 9.3. taulan ageri dira, hasierako esperimentuaren emaitzekin batera.

Azken esperimentu honetan, espero bezala, PER eta Zehaztasun emaitza hobeak lortu dira isiltasun GMMarekin batera hizketa GMMa erabiliz behaketa-egiantzak sortu eta MNSaren bidez MLPa elikatzeko. Emaitza horiek eta 9.1. taulakoak alderatzen baditugu, ikusten dugu ezen *online* emaitzarik onenak isiltasun GMMa eta hizketa GMMa erabiliz MNSan oinarritutako metodoarekin lortzen direla. PER emaitzak hobeak dira Zehaztasun balioak baino, eta emaitzak hobeak dira, gainera, hurbileko kanalerako, mahai gaineko kanalerako baino. Laburbilduz, MNSan oinarritutako normalizazioa alternatiba egokia da kapitulu honetan aurkeztutako *online* normalizazio-metodo guztietarako. Metodo berri hau ondo dabil seinale garbiekin, sortzen duen errorea ez baita oso handia. Seinale zaratatsuagoetarako, ordea, handiagoa da errorea.

Table 9.3: MNSan oinarritutako (*online*) CMVN inplementazioaren emaitzak, hizketa GMMa erantsita: *hurbileko* eta *mahai gaineko* seinaleen PERak eta Zehaztasunak, *offline* balioekin alderatuta.

	PER (%)		
	Offline	*Online*	
	Uneko datuak	MNSan oinarritua (isilt. GMM)	MNSan oinarritua (isilt. + hizk. GMM)
Hurbilekoa	12.45	15.51	13.18
Mahai gainekoa	21.87	29.29	25.28

	Zehaztasuna (%)		
	Offline	*Online*	
	Uneko datuak	MNSan oinarritua (isilt. GMM)	MNSan oinarritua (isilt. + hizk. GMM)
Hurbilekoa	74.24	65.50	71.36
Mahai gainekoa	60.12	45.45	50.22

9.5 Konklusioak

Kapitulu honetan, *online* CMVNari buruz jardun dugu. CMVNarekin, ezagutze-emaitzarik onenak lortzen dira MFCCen batezbestekoak eta bariantzak uneko seinale osoa erabiliz estimatzen direnean (*offline* metodoa). *Online* CMVNak badu desabantaila bat: normalizazioa seinale osoa iritsi aurretik aplikatu behar denez, beste metodo bati jarraituz estimatu behar dira batezbestekoak eta bariantzak. Gaur egun ere erronka handia da gai hau.

Inplementazio klasikoak hiru modutara sailkatu daitezke: iraganeko datuak erabiliz, non batezbestekoak eta bariantzak aurretiaz jasotako seinaleak erabiliz estimatzen baitira; *ikuspegi segmentala*, non CMVN leiho lerrakor batean aplikatuz normalizatzen baitira MFCCak; eta *ikuspegi errekurtsiboa*, non batezbesteko eta bariantza bektoreak uneko esakuntzako lehen D bilbeak erabiliz hasieratzen baitira eta, ondoren, errekurtsiboki eguneratzen baitira bilbe berriak iritsi ahala.

hiru ikuspegi horiek kontsideratuz, lau inplementazio testatu dira:

- **Iraganeko datuak erabiltzea**: Seinale osoak normalizatzeko, *Basque Speecon-like* datu-baseko hurbileko kanaletik erauzitako batezbestekoak eta bariantzak erabiltzen dira.

- **Iraganeko datuak eta eguneratze errekurtsiboa**: Batezbestekoen eta bariantzen hasierako balio gisa, *Basque Speecon-like* datu-baseko hurbileko kanaletik erauzitako batezbestekoak eta bariantzak erabiltzen dira, eta, ondoren, datozen bilbeen balioez eguneratzen dira.
- **Hasierako aurrera begirakoa eta eguneratze errekurtsiboa**: Batezbestekoen eta bariantzen hasierako balioak kalkulatzeko, D bilbe erabiltzen dira, eta, ondoren, datozen bilbeen balioez eguneratzen dira.
- **Teknika hibridoa**: Batezbestekoen hasierako balioak kalkulatzeko, D bilbe erabiltzen dira, eta, ondoren, datozen bilbeen balioez eguneratzen dira; bariantzak, ordea, *Basque Speecon-like* datu-baseko hurbileko kanaletik erauzitako balioak erabiltzen dira, balio konstanteak.

Inplementazio horiek mendekotasun handia dute hizketak seinalean zehar duten banaketarekiko. Adibidez, askoz emaitza hobeak lortuko aurrera begirakoak isiltasun-nahiz hizketa-bilbeak baldin badauzka. Gainera, eguneratze-modua erabiltzen bada, isiltasun-segmentu edo hizketa-segmentu luzeek desorekatu egiten dituzte batezbestekoak eta bariantzak, eta horrek eragin handia du emaitzetan. Horrek guztiak esan nahi du inplementazio horiek ez direla egonkorrak.

Guk proposatu dugun MNSan oinarritutako *online* CMVNak arazo horiek konpontzen ditu. MNSan oinarritutako metodoak ez du atzerapenik, zeren ez baita aurreko eta geroagoko bilbeen araberakoa. MNSan oinarritutako CMVNa erabiliz lortutako emaitzak hobeak dira. Gainera, berdin dio hizketa nola dagoen banatuta seinalean zehar; hortaz, berdin jokatzen du seinalean egon litekeen edozein hizketa- eta isiltasun-banaketarako. Desabantailarik handiena da metodoak zehaztasuna galtzen duela seinale zaratatsuetarako. MNSan oinarritutako *online* CMVNak ikerketa gehiago behar du normalizazioa sendoa lortzeko; halere, oso erabilgarria izan daiteke.

CHAPTER 10

Fonemak puntuatzea: GOPetatik DNNetara

10.1 Sarrera

6. kapituluan, fonemak puntuatzeari buruzko hasierako esperimentuak deskribatu dira. Ebakera-egokitasuna (GOP, *Goodness Of Pronunciation*) izeneko puntuazioak hautatu dira egiaztapen-puntuazio gisa, eta ebakera okerreko fonemen GOP puntuazioak lortzeko bidea izan da erroreak artifizialki simulatzea, ebakera-hiztegian aldaketak txertatuz. Horren oinarrian bada ideia bat: hizkuntza jakin batean fonema bat beste fonema baten errealizazioetatik zenbat eta hurbilago egon, orduan eta okerrago ebakirik egongo da.

Atalaseak lortzeko, zuzen ebakitako fonemetarako eta oker ebakitako fonemetarako errore berdinaren tasak (EER, *Equal Error Rate*) kalkulatu dira (erroreak simulatzeko teknikaz). Hala, EER puntu bat kalkulatu da fonema talde bakoitzerako. Laborategiko tesiek oso emaitza onak zituzten, baina, ingurune errealista batean egindako esperimentuetan, ikusi digu sistemaren funtzionamenduak okerragora egiten duela. Berez, fonema jakin batzuk nekez gainditzen zuten atalasea. Halaber, hasierako eta amaierako fonemek ez dute gainerako fonemen jokaera bera, eta, hortaz, berariazko GOP banaketak behar lirateke haien GOP atalaseak ezartzeko. Litekeena da fonemak hitzean duen kokapena kontuan izan beharreko ezaugarria dela.

Fonema bat zuzen ala oker ebaki den erabakitzeko GOP atalaseen multzo bat erabiltzeko metodoa mugatu samarra da. Gaur egungo lanek fonema-iraupenak, ondorengo probabilitateak, GOPak eta abar baliatzen dituzte [65]. Gainera, ondoko fonemei buruzko informazioa ere baliagarria izan daiteke, zeren oker ebakitako fonema batek eragina izan dezake uneko fonema-deskodetzean ez ezik, ingurukoetan ere; ez bakarrik GOPetan, baita iraupenetan eta probabilitateetan ere.

Linguistikan, fonema bat honela defini daiteke: "hizkuntza batean oinarrizko unitate kontrastatzaile bat osatzen duten hizketa-soinuen multzoa. [...] Elementuek (alofonoek) ez dute kontrasterik egiten batak bestearekin hizkuntza jakin horretan" [211]. Hortaz, fonemek pare minimoak osatzen dituzte hizkuntza jakin batean, eta, horrenbestez, ondoriozta dezakegu fonema baten ebakera-zuzentasunaren kontzeptuak lotura zuzena duela gainerako fonemekin; alegia, uler litekeela, hizkuntza jakin batean, fonema baten

errealizazioen eta gainerako fonemen errealizazioen arteko distantzien baitan —N dimentsioko espazio batean—.

10.2 Oker ebakitako fonemaren kontzeptua

Hizkuntza jakin batean, kontsidera daiteke fonema bat oker ebakita dagoela bere errealizazioa nolabait hurbilago dagoenean beste fonema baten errealizazioetatik, bere ohiko errealizazioetatik baino. Hala, 10.1. irudian ikus daitekeenez, jotzen badugu fonema baten errealizazioak N dimentsioko espazio bateko puntuak direla, p_1 fonemaren errealizazio berri bat "zuzen ebakitakotzat" jo liteke, baldin eta p_1 fonemaren ohiko errealizazioetatik (edo zentroidetik) hurbilago kokaturik badago, gainerako fonemen errealizazioetatik baino. Baldin eta $p_1$en errealizazioa beste fonema baten errealizazioetatik (edo zentroidetik) hurbilago badago, "oker ebakitakotzat" jo liteke. Arestian azaldu denez, distantzia edo tarte horiek desberdinak izango dira hizkuntza batetik bestera, eta gainerako fonemen errealizazioetarako distantzien araberakoak dira.

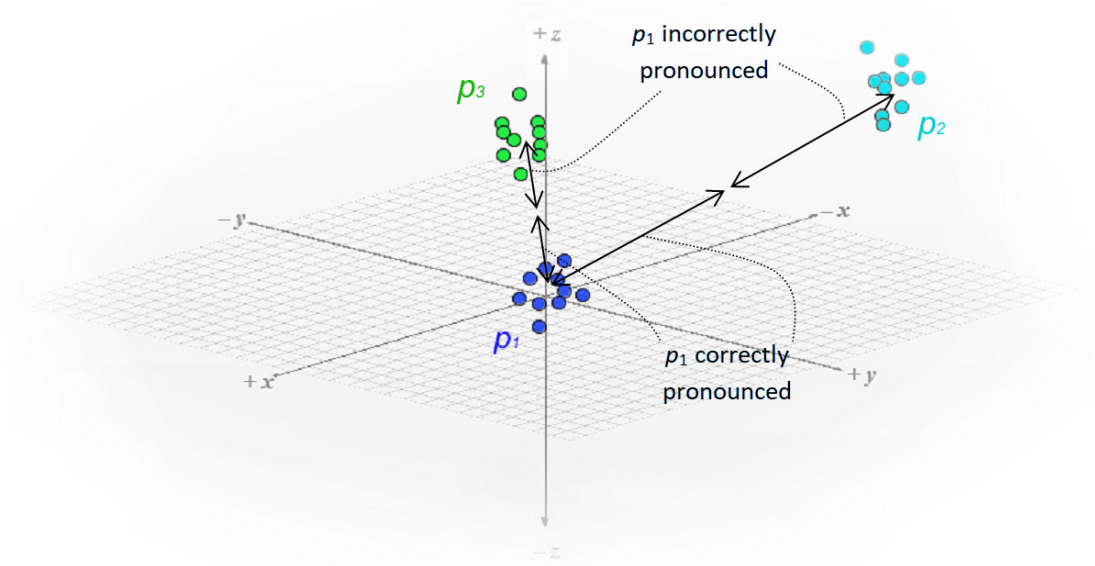


Figure 10.1: p_1 (urdin iluna), p_2 (urdin argia) eta p_3 (berdea) fonemen errealizazioak, 3 dimentsioko espazio batean.

Ikuspuntu horretatik, hizkuntzaren fonema-inbentarioaren araberakoa da zer neurritan dagoen fonema bat zuzen ala oker ebakita. Adibidez, hizkuntza batek 12 bokal ahokari baldin baditu (ikus SAMPA frantsesa[1]) eta beste batek 5 (euskarak, adibidez), "zuzen ebakitako" fonemen espazioa handiagoa izango da euskarazko bokal baten inguruan, frantseseko bokal baten inguruan baino. Labur esanda, hizkuntza jakin bateko fonema

1 https://www.phon.ucl.ac.uk/home/sampa/french.htm

baten ebakera-zuzentasuna hizkuntza horren inbentarioko gainerako fonemen eta haien arteko distantzien araberakoa da.

10.3 Entrenamendu-datuak

10.3.1 Entrenamenduko datu-multzoa

Entrenamenduko datuak hautatzean, irregulartasun eta akats kopururik baxuena duten hizketa-datuak hartu dira kontuan. Lan honetako lehen esperimentuetan bezala (ikus 6.2. atala), *Basque Speecon-like* datu-baseko *Train* blokeko *R+M25* azpimultzoa erabili da (*R* azpimultzoa —atal irakurria— gehi *M*25 azpimultzoa —eskuz zuzendutako 25 saio—), zehazki, ekialdeko euskal hiztunei dagozkien saioak, entrenamendu-datu gisa (155 erabiltzailetik, 76); izan ere, eremu horretan desberdin ahoskatzen dira euskarazko hiru txistukari igurzkariak, alde batetik, eta hiru txistukari afrikatuak, bestetik.

Audio-multzo horretan, guztira 761 503 fonema-errealizazio daude. Haietariko erdiak (% 50.03) bokalak dira; bestalde, gutxien ageri den fonema-taldea sabaikarien taldea da (% 1.24). Hautatutako datuetako fonema kopuruen xehetasun gehiago 10.1. taulan jaso dira. Azpimarratzekoak dira sabaikarien (*L*, *jj* eta *gj*) eta afrikatuen (*tz*, *ts* and *tS*) kopuru txikiak. Horrek eragin handia izango du emaitzetan, eta nabariagoa izango da afrikatuen kasuan, zeren, eskuarki, talderik arazotsuenetariko bat baita euskara-ikasleentzat.

Table 10.1: Entrenamendu-datuetako fonemen kopuruak (%), talde fonetikoka.

	Fonema kopuruak (%)
Bokalak	50.03
Igurzkariak	6.89
Afrikatuak	2.09
Herskari ahostunak	9.14
Herskari ahoskabeak	12.72
Sudurkariak	7.62
Sabaikariak	1.24
Likidoak	10.27

10.3.2 Oker ebakitako fonemak lortzea

Aurreko atalean deskribatu dugun oker ebakitako fonemaren kontzeptua kontuan hartuta, jo dezakegu ezen fonema bat erabat oker ebakitzea, hizkuntza jakin batean, hizkuntza horretako beste fonema bat ebakitzea dela. Errore simulatuen metodoa (6.2. atalean deskribatua) bat dator ideia horrekin, zeren sistemaren ebakera-hiztegian fonemak bata bestearekin ordezkatuz lortzen baita informazioa (kasu honetan, GOPak). Hala, ezagutzaileak, lerrokatze behartu moduan, ordezkatutako fonemaren HMMaz kalkulatu-

tako GOPak lortzen ditu, jatorrizko fonema bezala ahoskatzen ari dela kontsideratuz (oker ebakitzearen simulazioa). Kontuan izatekoa da fonemak talde akustiko bereko fonemekin soilik ordezkatzen direla; hala, kontserbadoreagoak dira emaitzak.

Esperimentu honetan, errepikatu egin dugu erroreak simulatzeko prozesua, baina audio-datu kopuru handiagoa erabiliz: fonema bakoitza talde bereko gainerako fonemekin ordezkatu eta banan-banan prozesatu da. Hala, informazio kopuru handiagoa lortu da errore simulatuetarako, hain zuzen 2 591 755 fonemaren informazioa. Zuzen ebakitako eta oker ebakitako datuen kopuruak orekatzeko, baztertu egin dira errore simulatu asko. Baldin eta erroreak simulatuz lortu diren fonemen kopurua zuzen ebakitakoena baino handiagoa bazen, oker ebakitako fonemak baztertuz joan gara, zuzen ebakitako fonemen kopuruarekin berdindu arte. Aitzitik, oker ebakitako fonemen kopurua zuzen ebakitako fonemen kopurua baino txikiagoa bazen, beren horretan utzi dira bi kopuruak. Azken kopuruak 10.2. taulan ageri dira, fonemen kokapenaren arabera sailkaturik.

Table 10.2: Entrenamendu-datuetako fonemen kopuru osoa (C: zuzen ebakitakoak; X: oker ebakitakoak, errore simulatuak), esakuntzan duen kokapenaren arabera eta talde fonetikoka.

	Ezk. fonemak		Erd. fonemak		Esk. fonemak		Bakarrak	
	C	X	C	X	C	X	C	X
Bokalak	26 812	26 812	308 982	308 982	44 742	44 742	499	
Igurzkariak	6 995	6 995	42 916	42 916	1 147	1 147	30	
Afrikatuak	367	157	14 911	9 534	684	684	0	
Herskari ahostunak	14 835	10 495	54 609	54 609	126	114	0	
Herskari ahoskabeak	6 549	6 549	84 618	79 736	5 618	3 725	0	
Sudurkariak	3 732	2 071	38 657	38 657	3 090	3 090	0	
Sabaikariak	529	529	8 000	6 492	305	305	0	
Likidoak	1 609	1 609	73 747	73 747	930	930	0	

10.3.3 Datuen analisia

6.2. atalean, zuzen eta oker ebakitako fonemetatik GOP puntuazioak erauzten dira, eta GMMak entrenatzen dira fonema talde bakoitzerako. Esperimentu honetan, iraupenak eta egiantzak ere erauzi dira. Gainera, inguruko fonemen GOPak, iraupenak eta egiantzak ere hartu dira kontuan, ikusteko ea haiek ere informazio erabilgarria izan dezaketen.

Atal honetan, GOPen, iraupenen eta log-egiantzen zenbait histograma erakusten ditugu, bai aztergai den fonemarako, bai haren aurreko zein atzeko fonemetarako. Hala, elementu bakoitzak egiten duen ekarpenari buruzko ideia orokor bat osatu ahalko dugu.

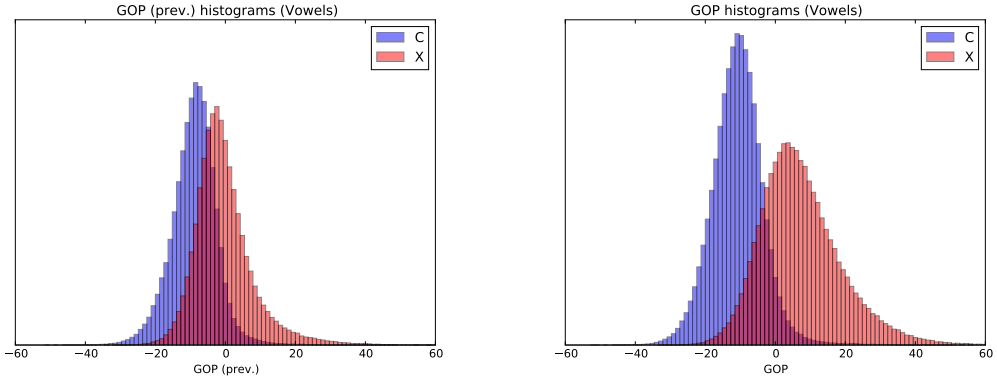


Figure 10.2: Zuzen (C) eta oker (X) ebakitako bokaletatik erauzitako GOP histogramak (eskuinean) eta aurreko fonemaren GOP histogramak (ezkerrean).

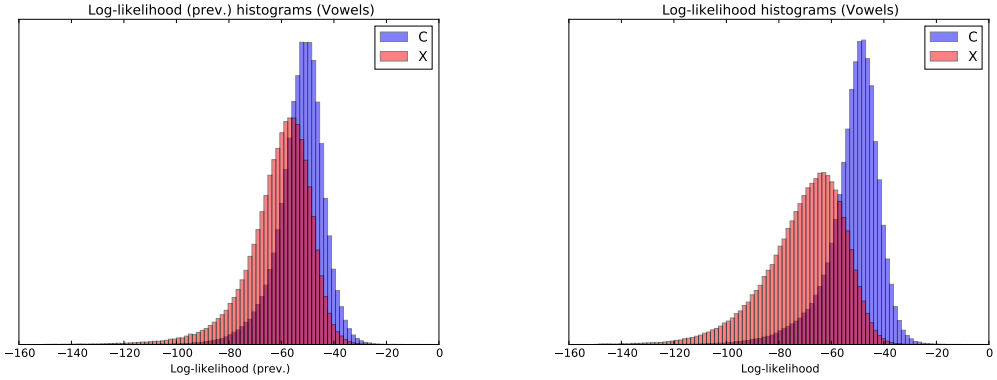


Figure 10.3: Zuzen (C) eta oker (X) ebakitako bokaletatik erauzitako log-egiantzen histogramak (eskuinean) eta aurreko fonemaren log-egiantzaren histogramak (ezkerrean).

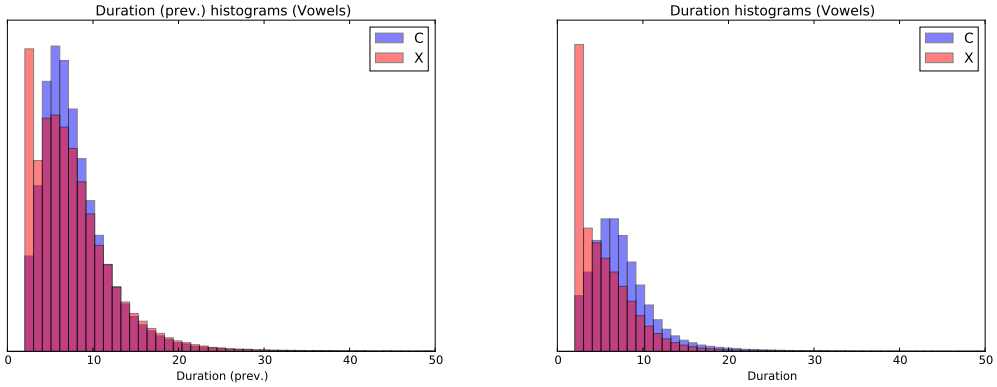


Figure 10.4: Zuzen (C) eta oker (X) ebakitako bokaletatik erauzitako iraupenen histogramak (eskuinean) eta aurreko fonemaren iraupenen histogramak (ezkerrean).

Ikus ditzagun kasurik onena eta okerrena: kasurik onena bokalek osatzen dute, bereziki tartekoek. 10.2. irudian, bokalen GOP histogramak ageri dira, unekoarenak (eskuinean) eta haren aurrekoarenak (ezkerrean). Uneko fonemaren GOP banaketek nahiko bereiziak diruditen arren, aurreko fonemaren banaketak aski gainjarrita daude. Log-egiantzei dagokienez (ikus 10.3. irudia), GOPentzat atera dugun konklusio bera atera dezakegu. Iraupenei dagokienez, berriz, bi banaketak daude oso gainjarriak, baina are gehiago aurreko fonemarako. Laburbilduz, dirudienez aurreko fonemak ez du informazio bereizlerik ematen, ezta iraupenak ere.

Bestalde, kasurik okerrenak sabaikariek, afrikatuek eta, maila xumeago batean, igurzkariek osatzen dituzte. Sabaikariei dagokienez, histograma guztiak ia erabat gainjarririk daude (ikus 10.5. irudia GOPentzat and 10.6. irudia log-egiantzentzat). Fonema horien arteko nahastea da horren zergatia (ikus azalpena 5.2.1. atalean). Gainera, fonema talde horrek badu eragozpen bat: *Basque Speecon-like* datu-baseko instantzia

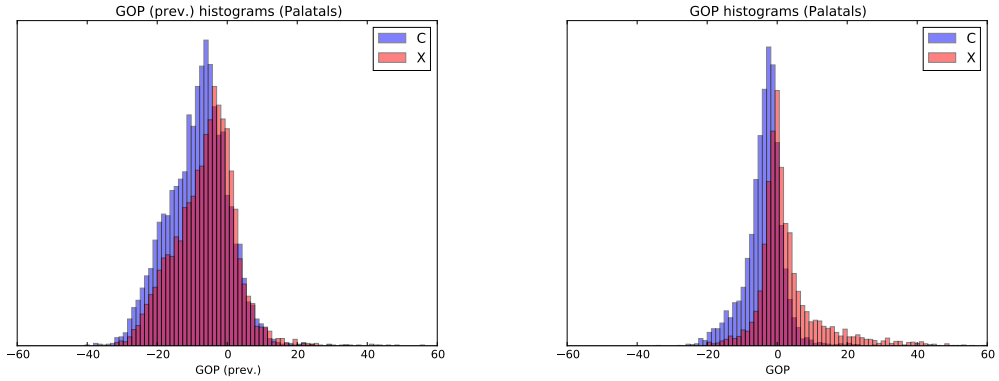


Figure 10.5: GOP histogramak (eskuinean) eta aurreko fonemaren GOP histogramak (ezkerrean), zuzen (C) eta oker (X) ebakitako sabaikarietarako.

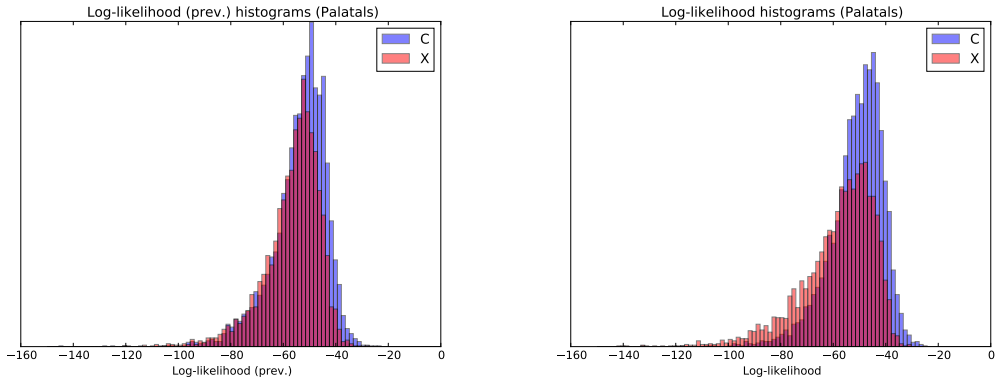


Figure 10.6: Log-egiantzak (eskuinean) eta aurreko fonemaren log-egiantzak (ezkerrean), zuzen (C) eta oker (X) ebakitako sabaikarietarako.

kopururik baxuena du.

Afrikatuen kasuan (10.7. irudia, eskuinean), histogramak aski gainjarririk daude (azpimarratzekoa da ezen bi banaketak erabat gainjartzen direla ekialdeko jatorrizko hiztunak erabili beharrean ez-jatorrizko hiztunak ere erabilita). Igurzkarietarako, histogramak pittin bat bananduago daude (10.7. irudia, ezkerrean).

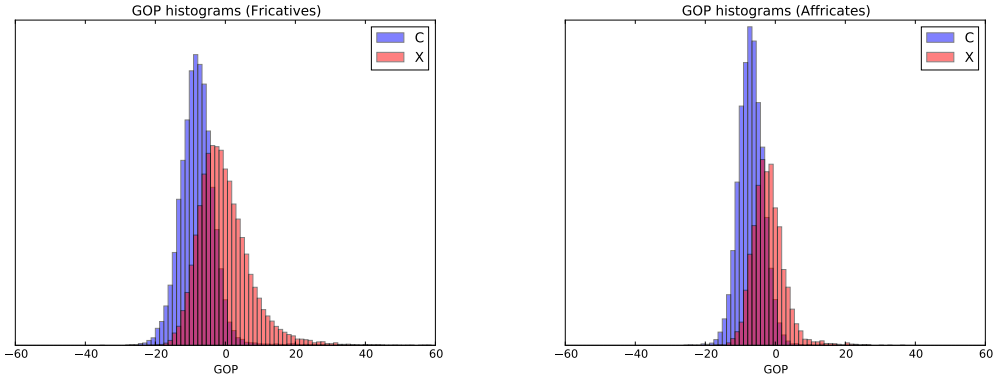


Figure 10.7: Zuzen (C) eta oker (X) ebakitako igurzkarietatik (ezkerrean) eta afrikatuetatik (eskuinean) erauzitako GOP histogramak.

Gainerako taldeei dagokienez (heskari ahostunak, herskari ahoskabeak, sudurkariak eta likidoak), GOP histogramek nahiko erabilgarriak dirudite desberdintzeko. Log-egiantzen histogramek gainjarriagoak daude. Adibide gisa, 10.8. irudian, herskari ahostunen GOPen eta log-egiantzen histogramak ageri dira.

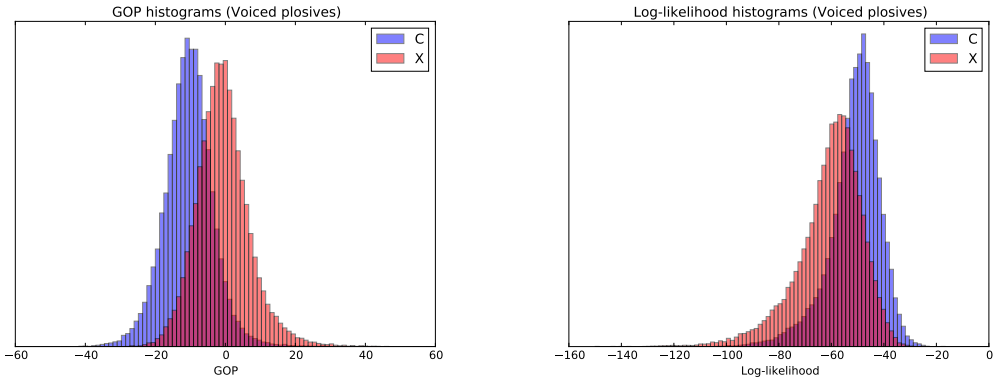


Figure 10.8: Zuzen (C) eta oker (X) ebakitako herskari ahostunetatik erauzitako GOP histogramak (ezkerrean) eta log-egiantzen histogramak (eskuinean).

Histograma guztien zerrenda bat ageri da A. eranskinean, fonema-taldearen eta hitzean duen kokapenaren arabera sailkaturik.

10.4 Erabakia hartzea: Neurona Sareak

[36]ko konklusioetan azaltzen denez, non ebakera-akatsei automatikoki detektatzeko metodoen errepaso bat egiten baita, bete gabe dugun erronkarik handienetako bat da gaur egungo sistemak integratu eta $L1$ekiko independentea den sistema bat lortzea edo, gutxienez, beste $L1$ baterako erraz konfigura daitekeena, ez-jatorrizko hizketa duen eta eskuz etiketatu behar den datu-base bat erabili beharrik gabe. Ideia horri jarraiki, neurona-sareak (NN, *Neural Networks*) erabiliko ditugu, ikusteko ea aukera badagoen eredu konposatu orokor bat sortzeko. Parametro bakoitzak (GOPek, iraupenek eta egiantzek) NN ereduan duen eragina ere ikusiko dugu hala.

10.4.1 Entrenamenduko parametro multzoak

Geruza anitzeko hainbat pertzeptroi (MLP, *Multi-Layer Perceptron*) entrenatu dira ikusteko ea baliagarriak diren fonema bat zuzen ala oker ebakita dagoen erabakitzeko. Lehendabizi, MLP bakar bat entrenatu da fonema guztietarako, sarrerako geruzan fonema bakoitzaren identitatea adierazita. Ondoren, MLP desberdinak entrenatu dira fonema talde bakoitzerako.

Gainera, parametro multzo desberdinak erabili dira MLPak entrenatzeko:

a) GOPak, iraupenak and log-egiantzak + aurreko eta ondorengo fonemen GOPak, iraupenak eta log-egiantzak..

b) GOPak, iraupenak and log-egiantzak.

c) GOPak eta log-egiantzak.

d) GOPak.

e) Aurreko, uneko eta ondorengo fonemen GOPak.

MLPak TensorFlow™ erabiliz sortu dira, prozesaketa numerikorako software-liburutegi bat, kode irekikoa, datu-fluxuen grafoak erabiltzen dituena eta prozesaketa CPU edo GPU batera edo gehiagotara bideratzeko arkitektura malgua duena. TensorFlow, jatorrian, *Googleko Machine Intelligence* ikerketa-erakundeko *Google Brain* taldean lanean zihardutenikertzaileek eta ingeniariek garatu zuten, ikasketa automatikoaren eta neurona-sare sakonen ikerketak bideratzeko. Gaur egun, konfigurazio desberdineko NNak entrenatzeko eta testatzeko erabiltzen da, batez ere.

10.4.2 Testeko datuak

Aurreko atalean azaldutako eredu desberdinak testatzeko, azken urteotan gure *online*ko CAPT[1] demoa erabilita noizbehinkako erabiltzaileek grabatutako audio-seinaleak erabili ditugu. 30 pertsonak utzi dituzte haien grabazioak gure zerbitzarian, eta ez dago inolako informaziorik erabiltzaileari buruz. Antza, euskara edo espainiera dira erabiltzaile guztien jatorrizko hizkuntzak. 24 gizonezko (56 fitxategi) eta 6 emakumezko (22 fitxategi) daude,

1 https://aholab.ehu.es/users/igor/CAPT

eta batez bestez 2.6 fitxategi daude hizlariko. Era guztietako hondo-zaratak, SNRak eta grabatzeko gailuak ageri dira grabazioetan.

Fitxategiak eskuz etiketatu dira. Guztira, 1269 fonema-errealizazio daude, eta haietariko 29 soilik daude oker ebakita. Datu errealek badu desabantaila bat: oker ebakitako askoz fonema gutxiago daude zuzen ebakitakoak baino, eta horrek desorekatu egiten du emaitzen analisia. Kontuan izatekoa da erabaki bitarra egin dela hemen; hortaz, fonema bat zuzen ebakita ala oker ebakita baino ezin da egon. Oker ebakitako fonemen kopurua 10.3. taulan ageri da, taldeka. Ia % 80 txistukariak dira, bai igurzkariak (z, s and S), bai afrikatuak (tz, ts and tS), espero zenez; izan ere, erabiltzaileen jatorrizko hizkuntza euskara edo espainiera da. Bestalde, herskari bat ere ez da oker ahoskatu.

Table 10.3: Testeko datuetan dagoen oker eta zuzen ebakitako fonemen kopurua.

	Oker	Zuzen
Bokalak	2	633
Igurzkariak (txistukariak)	13	125
Afrikatuak	10	83
Herskari ahostunak	0	75
Herskari ahoskabeak	0	103
Sudurkariak	2	113
Sabaikariak	1	14
Likidoak	1	94

10.4.3 Emaitzak

Emaitzak lortzeko erabili dugun metrika puntuatze-zehaztasuna da (SA, *Scoring Accuracy*). (10.1) ekuazioan ageri den bezala kalkulatzen da SA,

$$SA\,(\%) = \left(\frac{CA + CR}{CA + CR + FA + FR}\right) \cdot 100 \tag{10.1}$$

non: CA (*Correctly Accepted*): Zuzen onartutako unitateak; CR (*Correctly Rejected*): Zuzen baztertutako unitateak; FA (*Falsely Accepted*): Oker onartutako unitateak; FR (*Falsely Rejected*): Oker baztertutako unitateak.

10.4. taulan, test bakoitzean MLP bakoitzaz lortu diren SA emaitzak ageri dira. Taula horretako MLPak 64 nodoko ezkutuko geruza erabiliz entrenatu dira. 10.5. taulan, emaitza berak ageri dira, baina MLParen ezkutuko geruzan 6 nodo erabiliz.

Table 10.4: Parametro multzo desberdin batez eta 64 nodoko ezkutuko geruza erabiliz entrenatutako MLP bakoitzaz test bakoitzean lortutako puntuazio-zehaztasunak (SA).

	Testak (SA)				
MLPak	a	b	c	d	e
Eredu bakarra	57.76	64.46	62.81	**65.09**	58.79
Bokalak	63.62	69.61	69.76	**74.02**	65.83
Igurzkariak	50.00	**62.32**	55.05	60.14	49.28
Afrikatuak	35.48	**45.16**	40.86	40.86	37.63
Herskari ahostunak	50.67	52.00	54.67	**61.33**	49.33
Herskari ahoskabeak	42.72	46.60	42.72	**51.46**	43.69
Sudurkariak	65.22	74.78	**80.87**	77.39	**80.87**
Sabaikariak	60.00	60.00	**66.67**	**66.67**	53.33
Likidoak	67.37	**70.53**	65.26	69.47	68.42

Table 10.5: Parametro multzo desberdin batez eta 6 nodoko ezkutuko geruza erabiliz entrenatutako MLP bakoitzaz test bakoitzean lortutako puntuazio-zehaztasunak (SA).

	Testak (SA)				
MLPak	a	b	c	d	e
Eredu bakarra	60.91	65.41	59.47	**67.13**	59.73
Bokalak	65.83	70.71	69.67	**82.36**	65.35
Igurzkariak	52.90	68.12	60.87	**69.57**	50.00
Afrikatuak	38.71	43.01	39.78	**52.69**	36.56
Herskari ahostunak	49.33	54.67	57.33	**70.67**	49.33
Herskari ahoskabeak	44.66	40.78	42.72	**52.43**	42.72
Sudurkariak	65.22	79.13	79.13	78.26	**80.00**
Sabaikariak	46.67	66.67	66.67	**73.33**	66.67
Likidoak	64.21	67.37	65.26	**69.47**	61.05

Emaitzarik onenak ezkutuko geruzan 6 nodo erabiliz eta uneko GOPekin soilik entrenatutako MLPa erabiliz lortu dira. 6 nodoko testean, emaitzarik onenak, ia guztiak, d testean (uneko GOPak erabiliz) lortu dira; horrek esan nahi du ezen, oro har, gainerako parametroek zarata txertatzen dutela. Sudurkarien kasuan, aurreko eta ondorengo GOPek hobetu egiten dituzte emaitzak, baina ez gainerako kasuetan.

Ezkutuko geruzan 64 nodo daudenean, emaitzak banatuago daude: emaitza onenetariko 4 d testean lortzen dira, baina b-n 3, c-n 2 eta e-n 1 ere lortzen dira. Kasu honetan, emaitzak interpretatzean, esan dezakegu parametro multzo bakoitzak bere multzo optimoa duela, eta ezkutuko geruzako 64 nodoak gai direla desberdintasun horiek modelatzeko. Hala ere, emaitzak baxuagoak dira 6 nodorekin lortutakoak baino, sudurkarietarako, sabaikarietarako eta likidoetarako izan ezik. Beste test bat egin da ikusteko ea onuragarria litzatekeen ezkutuko geruzan nodo gehiago jartzea 128 nodoko ezkutuko geruza duen eredua entrenatuz, baina emaitzak okerragoak izan dira.

Azpimarratzekoa da SA balioak oso hurbil daudela CA balioetatik, oker eta zuzen ebakitako fonema kopuruen arteko desoreka dela eta. Horrenbestez, emaitzen jokaeraren irudi xeheagoa izateko, test onenaren taula jarri dugu (d testa, 6 nodorekin), non SAz gainera, CA, CR, FA eta FR ageri baitira (%-tan) (kontuan izan $CA + FR = \%\ 100$; eta $CR + FA = \%\ 100$).

Table 10.6: Ezkutuko geruzan 6 nodo erabiliz entrenatutako MLParen d testean lortutako SA, CA, CR, FA eta FR.

	SA	CA (%)	CR (%)	FA (%)	FR (%)
Bokalak	82.36	82.31	100.00	0.00	17.69
Igurzkariak	69.57	68.00	84.62	15.38	32.00
Afrikatuak	52.69	49.40	80.00	20.00	50.60
Herskari ahostunak	70.67	70.67	—	—	29.33
Herskari ahoskabeak	52.43	52.43	—	—	47.57
Sudurkariak	78.26	79.65	0.00	100.00	20.35
Sabaikariak	73.33	78.57	0.00	100.00	21.43
Likidoak	69.47	69.15	100.00	0.00	30.85

Taula hori kontu handiz aztertu behar da; izan ere, zuzen eta oker ebakitako elementuen kopurua oso desorekatuta dago. Esate baterako, ez dago oker ebakitako fonemarik herskarien artean, eta, beraz, berdinak dira SA eta CA, eta horrek ez digu adierazten nola jokatzen duen fonema zuzenen eta okerren sailkapenak. Bestalde, oker ebakitako oso fonema gutxi dauzkagu sudurkarietarako (2), sabaikarietarako (1) eta likidoetarako (1), eta, horren ondorioz, fonema horietako bat oker sailkatzeak % 100eko CRa edo % 100eko FA eragin lezake.

Bestalde, emaitzak bat datoz 10.3.2. ataleko irudiekin. Bokalen taldean bereizten dira ondoen zuzen eta oker ebakitako instantziak. Talde horretan lortzen da CArik onena, baita % 100eko CRa ere (nahiz eta oker ebakitako bi instantzia besterik ez egon). Sabaikarien, afrikatuen eta igurzkarien taldeak dira banaketa gainjarrienak dituzten

taldeak, eta, beraz, espero bezala, emaitzak ez dira oso onak. Emaitzarik harrigarriena herskari ahoskabeen CA da; izan ere, parametroen histogramari begiratzen badiogu, emaitza hobea espero genuen fonema-talde horretarako.

10.4.4 Konklusioak

Atal honetan, MLP desberdinak entrenatu dira parametro multzo desberdinak erabiliz, ikusteko ea parametro bat besteak baino eraginkorragoa den fonema bat zuzen ala oker ebaki den erabakitzeko unean. Oro har, MLParen ezkutuko geruzan 6 nodo erabilita, GOP puntuazioa da parametrorik diskriminatzaileena. Bestalde, ez dirudi aurreko eta ondorengo fonemek informazio erabilgarria ematen dutenik, ez GOParen bidez, ez gainerako parametroen bidez.

Ezkutuko geruzan, ordea, 64 nodo erabilita, taldekako emaitzarik onenak ez dira guztiak GOP puntuazioekin (d taldea) lortzen. Adibidez, likidoek b taldean lortzen dute guztizko emaitzarik onena; sudurkariek, berriz, c eta e taldeetan lortzen dute guztizko emaitzarik onena. Espero bazitekeen ere emaitzetarako onuragarria izan zitekeela MLPko ezkutuko geruzan nodo gehiago jartzea (baita geruza gehiago jartzea ere) ikusi dugu sistemak hobeto orokortzen duela nodo gutxi erabiliz.

10.5 Konklusioak

Kapitulu honetan:

- Zuzen eta oker ebakitako fonemen definizio bat eman dugu, seinale akustikoaren prozesamenduaren ikuspuntu praktiko batetik.
- Hainbat parametro aztertu ditugu, ikusteko haietariko zeinek ematen duen informazio erabilgarria, sailkatzaile gisa erabiliko dugun neurona-sarea (kasu honetan MLPa) entrenatzeko unean.
- Ikusi dugu GOP puntuazioak direla —atal honetan aztertutako parametroen artean— parametrorik eraginkorrenak. Gainerako parametroek zarata eransten dute hein batean edo bestean.
- Oro har, badirudi hobe dela MLParen ezkutuko geruzan 6 nodo erabiltzea, 64 baino.
- Parametrokako emaitzak nahiko koherenteak dira zuzen eta oker ebakitako fonemen histogramekin, herskari ahoskabeen kasuan izan ezik, espero baino emaitza baxuagoa baitu.
- Emaitza horiek sailkapen bitarra behar den kasurako lortu dira. Tarteko kategoria bat erabiliz gero, erabilgarriagoak lirateke emaitzak.

PART IV

Laburpena eta etorkizuneko lana

CHAPTER 11

Laburpena eta etorkizuneko lana

11.1 Tesiaren ekarpenak

Tesi honetan, ASRan oinarritutako euskararako bi OBHI aplikazio aztertu ditu: OBEL eta HHEE. Horretarako, ASR datu-base akustiko estandar bat erabili da, une honetan ez baitago euskarazko datu-baserik OBEL sistemak garatzeko balio duenik. Bi estrategiak fonemak egiaztatzeko tekniketan dute oinarria, eta horrek esan nahi du sistemak automatikoki erabaki behar duela fonema bat zuzen ala oker ebakita dagoen. Teknika berri bat proposatu dugu gai horri aurre egiteko: oker ebakitako fonemen populazioa lortzeko, errore lokalizatuak simulatu dira (errore artifizialak), eta, hala, zuzen ebakitako nahiz oker ebakitako fonemen GOP banaketak lortu ditugu. Horrekin, GOP banaketa pare bakoitzerako EER puntuak lor daitezke, atalase gisa erabil daitezkeenak. Puntu horiek doitu ere egin daitezke, sistema ikaslearen mailara doitu beharra egonez gero.

HMM desberdinak erabiliz lortu dira GOP atalaseak: OBELen, *Basque Speecon-like* datu-baseko jatorrizko hizlariak soilik erabili dira HMMak entrenatzeko, erreferentziazko fonema-errealizazioak behar baitira ikasleek ahoskatutako fonemak alderatzeko. HHEEn, ordea, euskaraz "behe-mailako" trebetasunak dituzten hizlariak ere erabili dira entrenamenduan; izan ere, sistemak ez-jatorrizko ikasleen hizketa prozesatu beharko du, gramatika-ariketak ahoz ebazten ari diren bitartean.

Online inplementazioa ere oso gai garrantzitsua izan da tesi honetan. Sistemaren helburua da web bidezko atzipen unibertsala lortzea, horrek bi abantaila nagusi baititu: batetik, prozesaketa intentsiboena zerbitzarian gertatzen da, eta, beraz, oso arinak lirateke erabiltzailearen gailuan exekutatuko liratekeen aplikazioak. Bestalde, inplementazioa HTML5en egiteak plataformarekiko independente bihurtzen du sistema.

Hala eta guztiz ere, zerbitzari-inplementazioak baditu zenbait desabantaila, erabiltzaile bakoitzak modu desberdin batean grabatuko baitu sarrerako audioa, alegia, gailu desberdin bat erabiliz eta ingurune desberdin batean. Arazo horri aurre egiteko, *online* CMVN berri bat aurkeztu dugu, tesi honetan proposatutako metodo batean oinarritua: *normalizazio anitzeko puntuatzea* (MNS, *Multi-Normalisation Scoring*). Proposatutako CMVN teknikaren bidez, seinale berri baten MFCCak bilbez bilbe normaliza daitezke,

atzerapenik gabe, eta emaitza lehiakorrak lortu dira. Gainera, *online* VAD sistema berri bat sortzeko ere baliatu dugu MNS metodoa, eta oso emaitza lehiakorrak lortu dira punta-puntako VAD sistemekin alderatuta, baina baldintza zaratatsuetan ere.

Bestalde, parametro multzo desberdinen azterketa bat egin da NNak entrenatzeko eta GOP atalaseen metodoa ordezkatzeko. Laburbilduz, GOP puntuazioak dira, kontuan hartutako parametroen artean, parametrorik eraginkorrenak. Gainerako parametroek zarata eransten dute hein batean edo bestean. Gainera, nahikoa dirudi ezkutuko geruzan nodo kopuru txiki bat erabiltzeak NNak entrenatzeko.

Tesi honen ekarpen orokorrak honela laburbil daitezke:

- ASRan oinarritutako OBHI sistemei buruzko literaturaren errepaso bat egin dugu. OBEL eta AGP aplikazioak dira, gaur egun, ezagunenak. GOP puntuazioak oso erabiliak dira, eta emaitza onak dituzte. Literaturako lan askotan, ikasleek oker ahoskatutako datuak erabiltzen dira eredu akustikoak moldatzeko, baina oso lan nekeza da OBEL aplikazioak sortzeko hain datu-base espezifikoak garatzea. Bestalde, badute desabantaila bat: ezin dira erabili $L1$ desberdin baterako tresnak garatzeko. Gaur egun, $L1$ekiko independentea izatea da OBEL sistemen diseinuen joera. Izan ere, nahiko erraza da $L1$-$L2$ pare jakin bateko zenbait akats ohiko lantzea, baina oraindik ere hor dirau sistema globalago baten beharrak. Gai horiek guztiak xehetasunez azaltzen dira 2. kapituluan.
- *Basque Speecon-like* datu-basea, tesi honetan erabilitakoa, sakonki aztertu eta moldatu egin da. Hasieran, datu-baseak audio-fitxategi guztiak zituen, baina transkripzio-fitxategien zati bat besterik ez. Beraz, transkripzio-lanak egin ziren tesi honen hastapenetan. Gainera, euskararako fonema-inbentarioa ere finkatu zen, eredu akustikoak entrenatzeko hitz-lexikoi egokia lortzeko. Aldaera dialektalak ere erantsi zitzaizkion lexikoiari, datu-baseko aldaera dialektal ugariak kontuan har zitezen. Datu-basearen bertsio optimizatua lagungarria izan da *SpeechTech4All* proiektua garatzeko, Espainiako Ekonomia eta Lehiakortasun Ministerioak finantzatutako proiektua, Espainiako hizkuntza ofizial guztietarako oinarrizko hizketa-teknologien ikerketa aurreratuan oinarritua. Lan hori guztia eta datu-basearen deskripzioa 3. kapituluan ageri dira.
- Hizketa ezagutzeko *AhoSR* sistema deskribatzen da 4. kapituluan. AhoSR-ren lehen bertsioa, sinplea baina egonkorra, 2012an egin zen, eta hitz-gramatika sinpleak maneiatzeko diseinaturik zegoen. Hala ere, garatuz joan da pixkanaka-pixkanaka, eta gaur egun hainbat ataza egin ditzake, hala nola ezagutze fonetikoa, hitz-gramatikan oinarritutako ezagutza eta hiztegi handiko hizketa-ezagutza (LVCSR, *Large Vocabulary Speech Recognition*). *Online* inplementazioaren bidez, AhoSR malguagoa da orain, eta zenbait proiektutan erabili da; adibidez, *Ber2Tek* proiektuan, Euskal Herriko hainbat agentez osatutako partzuergo batek bideratutako ikerketa-proiektu estrategikoan (azken demoan erabili zen). Gainera, tesi hau entregatzeko unean, Euskal Herrian hain prestigiotsua den *Elhuyar* Fundazioaren

online hiztegian inplementatuta dago (*Elhuyar*ren helburua da zientzia eta teknologia gizarteratzea eta euskararen aurrerapen teknologikoa ahalbidetzea); *AhoSR* hiztegiaren ahozko interfazea da, eta euskarazko hitzak eta terminoak ezagutzen ditu (une honetan hemen: https://hiztegiak.elhuyar.eus/).

- HMMen azterketa sakon bat egin da 5. kapituluan. Eredu akustiko horiek *Basque Speecon-like* datu-basea erabiliz entrenatuta daude, kontuan hartuta datu-baseak oraindik ere baduela atal bat ez dagoena behar bezala transkribatuta, alegia, bat-bateko hizketaren atala. Zuzeneko zenbait konklusio atera dira:
  - 32 gaussiar nahikoa dira datu-base honentzat, zeren, puntu horretatik aurrera, errore fonetikoaren tasak gora egiten du berriro ere.
  - Ebakera alternatibodun hiztegia, automatikoki sortua, ez da gai eredu akustiko hobeak sortzeko. Emaitzak, gainera, okerragoak dira eskuz zuzendutako transkripzioak erabili ezean, baina aldea txikiagoa da zuzendutako datuak erabiltzen badira. Horrek zera iradokitzen du: ebakera alternatiboak erabiltzearen onurak ikusteko, zuzendutako transkripzio gehiago behako liratekeela.
  - Entrenamendu-prozesuko etapa desberdinetan datu-multzo desberdinak erabiliz gero, hobetu egin daiteke emaitzen kalitatea. Hasieran eskuz zuzendutako atal bat eta azpimultzo irakurria eta, ondoren, datu-base osoa erabiliz, hobekuntzak lor daitezke.
  - Using CMVN better recognition results are obtained when analysing audio signals with mismatching channels, up to 30 % better. The recognition results testing audio signals recorded through the same audio channel are very similar to those obtained without cepstral normalisation.
  - CMVN erabiliz, ezagutze-emaitza hobeak lortzen dira kanal desberdinen bidez grabatutako audio-seinaleak analizatzean, % 30eraino hobeak. Oso antzekoak dira audio-kanal bereko seinaleak testatzen lortutako ezagutze-emaitzak eta normalizazio cepstralik gabe lortutakoak.
- 6. kapituluan, HHEEan oinarritutako AGP sistemaren lehen ebaluazioa deskribatzen da, AhoSR-ren lehen prototipoaz eginda dagoena. Han, GOP puntuazioen bi banaketa (zuzen eta oker ebakitako fonemei dagozkienak) lortu eta erabaki-atalase gisa banaketa bien EERa erabiltzearen estrategia berria testatu zen lehendabizikoz. Laborategiko ebaluazioek oso emaitza onak dituzte, baina inguru errealistago bateko testak behar dira. Ebaluazioa bi euskaltegi desberdinetan egin zen, behemailako euskara-ikasleen artean, eta % 89.89ko Zehaztasuna lortu zen emaitza gisa. Ikasleen artean sistemaren erabilgarritasunari buruz eta iritzi orokorrari buruz banatutako inkesta labur batek erakutsi zuen ezen erabiltzaileak funtzionamenduari buruz duen pertzepzioa pixka bat txikiagoa zela. Hala eta guztiz ere, emaitza itxaropentsuak lortu ziren.

- OBHI sistemen gaur egungo joera Internet bidezko bezero-zerbitzari arkitektura da. Audioa jasotzea plataformarekiko independentea izateak buruhauste ugari sortu ditu urte askoan, baina, gaur egun, azken urteotan hedatu den HTML5aren eta haren Audio APIaren bidez, ordenagailu batera konektatutako edozein mikrofonotako audioa atzi dezakete nabigatzaileek. HHEEan oinarritutako AGP atazetan, jaso ahal bidali behar da audioa. Horretarako, konexio bat behar da nabigatzailearen eta zerbitzariaren artean, eta hori, berez, ezinezkoa zen web teknologiaz. Hala ere, HTML5en *websocket*ak daude zehaztuta (web APIaren barnean), eta, haien bidez, *socket* moduko konexioak ezar daitezke nabigatzailearen eta zerbitzariaren artean. Bi API horiek konbinatuz, audioa *online* bidaltzeko sistema bat lor daiteke, HTML5 inplementaturik duen edozein nabigatzaile erabiliz.
- Hasierako sistemak zenbait hobekuntza izan ditu. Horietariko bat da VAD teknika berri bat, 8. kapituluan azaltzen dena. Oinarrian, tesi honetan proposatutako MNS metodoa du, zeinak modela daitezkeen patroiak sortzen laguntzen baitu. VAD sistema berri horrekin lortutako emaitza orokorrak oso lehiakorrak dira gaur egungo puntako beste sistema batzuekin alderatuta. Hizketa sailkatzeko erroreei dagokienez, testatutako VAD sistema onenaren antzeko emaitzak lortu dira; hala ere, isiltasuna sailkatzeko erroreak nabarmen hobeak dira gainerako sistemekin konparatuta. Are gehiago, emaitzak ez dira asko okerragotzen zaratez. VAD hori lehen aipatutako *Ber2Tek* proiektuaren azken demoan inplementatu zen, baita *Elhuyar*ren *online*ko hiztegian ere. MNSan oinarritutako *online* VADa azaltzen duen artikulua Q1 aldizkari batean argitaratu da (https://www.journals.elsevier.com/expert-systems-with-applications). VADaren *offline* bertsioa ere probatu daiteke hemen: https://aholab.ehu.es/users/igor/VAD/index.php.
- HHEEan oinarritutako AGP sistema webean inplementatzeko, MFCCak audioa iritsi ahala normalizatzeko metodo bat behar da. Literaturan, estrategia desberdinak erabili dira arazo horri aurre egiteko, eta haietariko batzuk testatu dira tesi honetan. Arazorik handiena da nola estimatu batezbestekoen eta bariantzen parametro cepstralen hasierako balioak, seinalea narriatu gabe. Emaitzarik onenak metodo hibrido batez lortu dira, hemen aurkeztutako metodo berri bat, datubasetik erauzitako bariantza-balio konstanteak eta lehen N bilbeetatik erauzi eta bilbez bilbe eguneratzen doazen batezbesteko-balioak darabiltzana. Gainera, *online* CMVN teknika berri eta sendoago bat ere proposatu da tesi honetan. Hori ere MNSan oinarrituta dago, eta, beraz, aurrez entrenatu daiteke. MNSan oinarritutako CMVNak ez du atzerapenik, ez baitu mendekotasunik aurreko eta ondorengo bilbeekin, eta emaitzak itxaropentsuak dira. Arazorik handiena da metodoak zehaztasuna galtzen duela seinale zaratatsuekin. Hori guztia 9. kapituluan deskribatuta dago. Une honetan, MNSan oinarritutako metodo deskribatzen duen artikulu bat prestatzen ari gara, argitaratzeko.
- Fonema-taldeen atalaseak ezartzeko, DNNak ere erabili dira. Lehendabizi, oker ebakitako fonemaren kontzeptua berrikusi dugu. Ideia horretan oinarrituta, zenbait

DNN entrenatu dira parametro multzo desberdinak erabiliz, parametro bakoitzaren eragina ikustearren. Esperimentuen emaitzek erakusten dute GOP puntuazioak direla parametrorik eraginkorrenak, aurreko, uneko eta ondorengo fonemen iraupenen eta log-egiantzen artean. Esperimentuen emaitzak koherenteak dira hasierako sistemaz lortutakoekin.

11.2 Emaitzak hedatzea

Argitalpenak

Ondorengo artikulu hau prestatzen ari gara:

"On-line Cepstral Mean and Variance Normalisation (CMVN) based on Multi Normalisation Scoring (MNS)": to be sent to *IET Electronics Letters* journal.

Ikerketa honi loturik, honako artikulu hauek argitaratu dira:
Aldizkariak:

1. **Igor Odriozola**, Inma Hernaez, Eva Navas: 'An on-line VAD based on Multi-Normalisation Scoring (MNS) of observation likelihoods'. *Expert Systems with Applications* (2018), vol. 110, pp. 52–61 (JCR Impact Factor: 3.768, Q1)
2. **Igor Odriozola**, Luis Serrano, Inma Hernaez, Eva Navas: 'The AhoSR Automatic Speech Recognition System'. *Advances in Speech and Language Technologies for Iberian Languages (Lecture Notes in Computer Science)* (2014), vol. 8854, pp. 279–288.
3. **Igor Odriozola**, Oliver Jokisch, Inma Hernaez, Rüdiger Hoffmann: 'Diseño y desarrollo de un sistema de evaluación automática de la pronunciación para el euskara'. *Procesamiento del Lenguaje Natural* (2012), vol. 49: pp. 101–108 (in Spanish).
4. **Igor Odriozola**, Eva Navas, Jon Sanchez, Inma Hernaez: 'Tratamiento léxico del euskara occidental basado en la división de radical y desinencia para reconocimiento de habla dialectal'. *Procesamiento del Lenguaje Natural* (2009), vol. 43: pp. 103–111 (in Spanish).
5. **Igor Odriozola**, Inma Hernaez, Eva Navas: 'Euskara eta Hizketa Teknologiak (Basque language and speech technologies)'. *BAT Journal of Sociolinguistics* (2008), vol. 66, pp. 125–133 (in Basque).

Tesi honetako ikerketarekin lotura zuzena duten kongresuetan egindako ekarpenak:

1. **Igor Odriozola**, Inma Hernaez, Eva Navas, Luis Serrano, Jon Sanchez: 'The observation likelihood of silence: analysis and prospects for VAD applications'. *Proc. of IberSPEECH (ISCA)* (2018), pp. 50–54, Barcelona (Spain).

2. **Igor Odriozola**, Inma Hernaez, M. Ines Torres, L. Javier Rodriguez-fuentes, Mikel Penagarikano, Eva Navas: 'Basque Speecon-like and Basque SpeechDat MDB-600: speech databases for the development of ASR technology for Basque'. *Proc. of International Conference on Language Resources and Evaluation (LREC) (ELRA)* (2014), pp. 2658—2665, Reykjavik (Iceland).
3. **Igor Odriozola**, Oliver Jokisch, Inma Hernaez, Rüdiger Hoffmann: 'A pronunciation tutoring system for Basque - First development steps'. *Elektronische Sprachsignalverarbeitung (ESSV)* (2012), pp. 101–108, Cottbus (Germany).
4. **Igor Odriozola**, Eva Navas, Inma Hernaez, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Daniel Erro: 'Using an ASR database to design a pronunciation evaluation system in Basque'. *International Conference on Language Resources and Evaluation (LREC) (ELRA)* (2012), pp. 4122–4126, Istanbul (Turkey).
5. **Igor Odriozola**, Inma Hernáez, Eva Navas: 'Design of a message verification tool to be implemented in CALL systems'. *Proc. of IberSPEECH (ISCA)* (2012), pp. 251–259, Madrid (Spain).
6. Igor Leturia, Arantza del Pozo, David Oyarzun, Urtza Iturraspe, Xabier Arregi, Kepa Sarasola, Arantza D. de Ilarraza, Eva Navas, **Igor Odriozola**, Iñaki Sainz: 'Web Communication Protocols for Coordinating the Modules of AnHitz, a Basque-Speaking Virtual 3D Expert on Science and Technology'. *Proc. of Web Services and Processing Pipelines in HLT workshop (LREC workshop) (ELRA)* (2010), pp. 60–67, Valletta (Malta).
7. Iker Luengo, Eva Navas, **Igor Odriozola**, Inma Hernaez, Iñaki Sainz, Daniel Erro: 'Modified LTSE-VAD Algorithm for Applications Requiring Reduced Silence Frame Misclassification'. *Proc. of International Conference on Language Resources and Evaluation (LREC) (ELRA)* (2010), pp. 1539–1544, Valletta (Malta).
8. Ibon Saratxaga, Inma Hernaez, **Igor Odriozola**, Eva Navas, Iker Luengo, Daniel Erro: 'Using Harmonic Phase Information to Improve ASR Rate'. *Interspeech (ISCA)*. (2010), pp. 1185–1188, Makuhari (Japan).
9. Igor Leturia, Arantza del Pozo, Kutz Arrieta, Urtza Iturraspe, Kepa Sarasola, Arantza D. de Ilarraza, Eva Navas, **Igor Odriozola**: 'Development and Evaluation of AnHitz, a Prototype of a Basque-Speaking Virtual 3D Expert on Science and Technology'. *International Multiconference on Computer Science and Information Technology (IMCSIT)*. (2009), pp. 235–242, Mragowo (Poland).
10. Gotzon Aurrekoetxea, Jon Sanchez, **Igor Odriozola**: 'EDAK: A corpus to analyse linguistic variations'. *Proc. of A Survey on Corpus-based Research / Panorama de investigaciones basadas en corpus (AELINCO)* (2009), pp. 489–503, Murcia (Spain).

Beste argitalpen batzuk

Hizketa-teknologiekin lotura duten eta parte hartu dudan beste lan batzuk:

1. Inma Hernaez, Eva Navas, **Igor Odriozola**, Kepa Sarasola, Arantza D. de Ilarraza, Igor Leturia, Beñat Oihartzabal, Jasone Salaberria: 'The Basque Language in the Digital Age – Euskara Aro Digitalean'. *META-NET White Paper Series* (2012).
2. Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernaez, Jon Sánchez, Ibon Saratxaga, **Igor Odriozola**: 'Versatile Speech Databases for High Quality Synthesis for Basque'. *Proc. of International Conference on Language Resources and Evaluation (LREC) (ELRA).* (2012) pp. 3308–3312, Istanbul (Turkey).
3. Ibon Saratxaga, Inma Hernaez, Eva Navas, Iñaki Sainz, Iker Luengo, Jon Sanchez, **Igor Odriozola**, Daniel Erro: 'AhoTransf: A Tool for Multiband Excitation Based Speech Analysis and Modification'. *Proc. of International Conference on Language Resources and Evaluation (LREC) (ELRA)* (2010), pp. 3732–3737, Valletta (Malta).
4. Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernaez, Jon Sanchez, Ibon Saratxaga, **Igor Odriozola**, Iker Luengo: 'Aholab Speech Synthesizers for Albayzin 2010'. *Proc. of VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop - FALA* (2010), pp. 343–347, Vigo (Spain).
5. Daniel Erro, Iñaki Sainz, Iker Luengo, **Igor Odriozola**, Jon Sanchez, Ibon Saratxaga, Eva Navas, Inma Hernaez: 'HMM-based Speech Synthesis in Basque Language using HTS'. *Proc. of VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop - FALA* (2010), pp. 67–70, Vigo (Spain).
6. Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernaez, Ibon Saratxaga, Iker Luengo, **Igor Odriozola**: 'The AHOLAB Blizzard Challenge 2009 Entry'. *Blizzard Challenge 2009 (http://festvox.org/blizzard/blizzard2009.html).* (2009), Edinburgh (UK).
7. Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, **Igor Odriozola**, J.J. Igarza, Inma Hernaez: 'Grabación de una Base de datos Bilingüe Euskera/Castellano para Verificación de Locutor'. *Proc. of V Jornadas en Tecnología del Habla* (2008), pp. 195–198, Bilbao (Basque Country).
8. Ibon Saratxaga, Eva Navas, Inma Hernaez, Jon Sanchez, Iker Luengo, **Igor Odriozola**, Eneritz de Bilbao: 'Evaluación Subjetiva de una Base de Datos de Habla Emocional para Euskera'. *Proc. of V Jornadas en Tecnología del Habla* (2008), pp. 191–194, Bilbao (Basque Country).
9. Iñaki Sainz, Inma Hernaez, Eva Navas, Jon Sanchez, Iker Luengo, Ibon Saratxaga, **Igor Odriozola**, Eneritz de Bilbao, Daniel Erro: 'Descripción del Conversor de Texto a Voz AhoTTS Presentado a la Evaluación Albayzin TTS

2008'. *Proc. of V Jornadas en Tecnología del Habla* (2008), pp. 96–99, Bilbao (Basque Country).

10. Iñaki Sainz, Ibon Saratxaga, Eva Navas, Inma Hernaez, Jon Sanchez, Iker Luengo, **Igor Odriozola**: 'Subjective Evaluation of an Emotional Speech Database for Basque'. *Proc. of International Conference on Language Resources and Evaluation (LREC) (ELRA)* (2008), paper 437, Marrakech (Morocco).
11. Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, **Igor Odriozola**, Inma Hernaez: 'Text independent speaker identification in multilingual environments'. *Proc. of International Conference on Language Resources and Evaluation (LREC) (ELRA)* (2008), paper 461, Marrakech (Morocco).
12. Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, **Igor Odriozola**, J.J. Igarza, Inma Hernaez: 'Building a Basque/Spanish bilingual database for speaker verification'. *Workshop Collaboration: interoperability between people in the creation of language resources for less-resourced languages (SALTMIL)* (2008), pp. 23–26, Marrakech (Morocco).
13. Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, **Igor Odriozola**, Inma Hernaez: 'Identificación de locutores en entornos multilingües'. *Proc. of IV Jornadas en Reconocimiento Biométrico de Personas* (2008) pp. 133-140, Valladolid (Spain).

11.3 Etorkizuneko lana

Lan hau amaituta, kontsideratzen dugu prest dagoela euskarazko OBEL sistema eta HHEEan oinarritutako AGP sistema urrutiko zerbitzari batean inplementatzeko teknologia. Lan honetan aurkeztutako sistemek beren hasierako helburuak bete dituzte, eta, gainera, soluzioak proposatu ditugu zerbitzari batean inplementatzean agertzen diren arazoak konpontzeko, hala nola *online*ko normalizazio cepstrala.

Garatu dugun teknologian oinarrituta, proiektu bat ari gara zehazten OBEL sistema euskara ikasteko HABEren (Helduen Alfabetatze eta Berreuskalduntzerako Erakundea) webgunean, *Ikasbil*en (www.ikasbil.eus), inplementatzeko. HABE Eusko Jaurlaritzako erakundea da, helduei euskara ikasten laguntzen diharduena, 107 euskaltegiren bidez eta 35 000 ikasle inguru (2016/2017 eskola-ikasturtean) izanik. Ikasleen % 12.24k autoikaskuntza hautatu du. Bestalde, Euskal Herritik kanpoko 3000 eta 4000 pertsona artean daude urtero HABEren bidez euskara ikasten dutenak, *Ikasbil*en bidez. Proiektuaren helburua da uneko OBEL sistema tresna gehigarri bat bezala inplementatzea, eta, pausuz pausu, behar diren hobekuntzak egitea OBEL sistema sendo bat lortu arte. Sistema horretan, fonemen errealizazioez gainera, prosodia ere ebaluatuko da.

Etorkizuneko lan behar-beharrezko bat da hizketa ezagutzeko sistemak neurona-sareak integratzea. Lehen urrats gisa, DNNz entrenatutako eredu akustikoak integratu beharko lirateke, ezagutze-emaitza hobeak lortzeko ez ezik, GOP eraginkorragoak entrenatzeko ere bai ([65]n azaltzen den moduan). Parametro esanguratsuak zein diren identifikatzea

eta ahoskatutako fonemak sailkatzeko DNNak erabiltzea ere etorkizunean aztertu beharreko ildoak dira. Horrekin guztiarekin, sistema sendoago, erabiltzeko errazago eta unibertsalago bat lortuko genuke.

APPENDIX A

AhoSR-ren parametro konfiguragarriak

Eranskin honetan, *AhoSR*-ren parametro garrantzitsuenak daude, baita haiei eman dakiekeen balioak ere. Parametro/balioa pareak konfigurazio-fitxategi batean ezar daitezke; hala, balio lehenetsiak gainidatziko lirateke aplikazioa kargatzen den unean. Izartxo (*) batekin markatutako balioek parametroaren balio lehenetsia adierazten dute.

Beste parametro batzuk, hala nola fitxategien kudeaketarekin edo datuak gordetzearekin lotuak, ez dira kontuan hartu, ez baitira esanguratsuak. Parametro horiek baliabide desberdinen kokalekuak adierazten dituzte, edo emaitzen, audio-fitxategien edo MFCC fitxategien irteera-datuen formatua.

A.1 Parametro orokorrak

Funtzionatzeko modu orokorrarekin eta audio-sarrerarekin lotutako parametroak.

Table A.1: Parametro konfiguragarri orokorrak *AhoSR*n

Parametroa	**Balioa**	**Deskripzioa**
OROKORRA		
OPERATION MODE	*RECOGNITION**	Performs the recognition decoding process.
	MFCC	Extracts MFCCs of the input signal and stores them in a file.
AUDIO MODE	*WAV_AUDIO**	Wav format files.
	DIRECT_AUDIO	Direct audio from microphone.
	SOCKET_RAW_-AUDIO	Raw PCM audio data through a socket connection.

A.2 Audio-sarrera

Sarrerako audio-seinalearen kalitatearekin lotutako parametroak.

Table A.2: Sarrerako audioaren kalitatearekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	**Balioa**	**Deskripzioa**
AUDIO SARRERA		
SAMPLE RATE	*16000**	Sample rate at which the incoming audio signal is processed.
BITS PER SAMPLE	*16**	Number of bits used to quantify audio samples.

A.3 MFCC erauzketako parametroak

MFCCen erauzketarekin eta ezaugarriekin lotutako parametroak.

Table A.3: MFCCen erauzketarekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	**Balioa**	**Deskripzioa**
MFCC PARAMETROAK		
FRAME RATE	*10**	The shift of the analysis window (in ms).
FRAME LENGTH	*25**	The length of the analysis window (in ms).
WINDOW TYPE	*HAMMING**	Hamming window.
	RECT	Rectangular window.
	BART	Bartlett window.
	HANNING	Hanning window.
	BLACK	Blackman window.
MIN. FREQ. LIMIT	*0**	Lower limit for frequency analysis.
MAX. FREQ. LIMIT	*0**	Upper limit for frequency analysis.
NUMBER OF FILTERS	*26**	Number of MEL filters for the frequency analysis.
SCALE	*MEL**	Mel scale.

hurrengo orrialdean darrai . . .

Parametroa	Balioa	Deskripzioa
MFCC PARAMETROAK		
	BARK	Bark scale.
NUMBER OF CEPSTRUM	*12**	Output number of Ceps coefficients.
C0	*1**	Includes 0th coefficient.
DELTA WINDOW	*2**	Window length for first derivatives ($X * 2 + 1$).
ACC WINDOW	*2**	Window length for second derivatives ($X * 2 + 1$).
LIFTERING	*22**	Liftering coefficient.
CMS	*false**	Apply CMS.
PREENF	*0.97**	Pre-emphasis coefficient.
HTK	*false**	To parametrise as in HTK.

A.4 Ezagutze-ataza

Erabili nahi den ezagutze-ataza motarekin lotutako parametroak.

Table A.4: Ezagutze-atazarekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	Balioa	Deskripzioa
EZAGUTZE ATAZA		
WORD NETWORK	*PHONETIC*	Phonetic recognition, using a structured triphone net.
	*WORD_GRAMMAR**	BNF and SLF grammar based recognition
	WORD_LOOP	Free word loop (choose LM for Continuous Speech Recognition).
	SENTENCE_VERIFICATION	Sentence verification (for CAPT purposes)
	ON-LINE_VERIFICATION	Sentence verification with word-by-word on-line feedback (for WWSV purposes).
	FORCED_ALIGNMENT	Forced alignment, both at word and phone level.
	MULTIPLE	Different grammars loaded and managed.

A.5 Hizkuntza Eredua

Hizkuntza-ereduen (LM, *Language Modelling*) konfigurazioarekin lotutako parametroak.

Table A.5: Hizkuntza Ereduekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	**Balioa**	**Deskripzioa**
HIZKUNTZ EREDUA		
LM	*false**	Use language model probabilities.
LM MODE	*3-GRAM**	Definition of the highest N-gram order.

A.6 Bilaketa-espazioaren antolaketa

Bilaketa-espazioko nodoen antolaketarekin lotutako parametroak.

Table A.6: Bilaketa-espazioaren antolaketarekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	**Balioa**	**Deskripzioa**
BILAKETA ESPAZIOA		
COMPRESSION TYPE	*NONE**	No compression is applied.
	PREFIX	Left-to-right node compression
	SUFFIX	PREFIX and right-to-left node compression.
UNIT EXPANSION	*true**	Word-end nodes are expanded to model coarticulation between words.
NON SPEECH EVENTS	*true**	Non speech events' HMMs are added in parallel to silence nodes.

A.7 HMMak

HMMen tipologiarekin lotutako parametroak.

Table A.7: HMMen tipologiarekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	**Balioa**	**Deskripzioa**
HMMak		
STATE NUMBER	*5**	State number of HMM (including end states).

hurrengo orrialdean darrai . . .

Parametroa	Balioa	Deskripzioa
HMMak		
GAUSSIAN NUMBER	*32**	Gaussian number of each GMM in each HMM state.

A.8 Inausketa

Inausketa-teknika desberdinekin lotutako parametroak.

Table A.8: Inausketarekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	Balioa	Deskripzioa
INAUSKETA		
PT	*0**	Pruning Threshold of histogram pruning.
MAX PATHS	*0**	Maximum number of active tokens permitted at each instant.
GAUSSIAN SELECTION	*0**	No Gaussian selection is applied.
	1	PDE (Partial Distance Elimination)
TOKEN NUMBER	*1**	Number of tokens that each state can hold.

A.9 CMVN

Parameters related to Cepstral Mean and Variance Normalization (CMVN) technique.

Batezbeste- eta bariantza-normalizazio cepstrala (CMVN) teknikarekin lotutako parametroak.

Table A.9: CMVNarekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	Balioa	Deskripzioa
CMVN		
CMVN	*0**	No normalization is applied.
	1	Mean normalization.
	2	Mean and variance normalization.
CMVN ONLINE	*false**	On-line calculation of means (and variances).
CMVN UPDATING	*false**	Recursive updating of means (and variances).

hurrengo orrialdean darrai . . .

Parametroa	Balioa	Deskripzioa
CMVN		
CMVN INIT LOOK-AHEAD	*25**	Number of initial frames to estimate the initial values of means (and variances).
CMVN INIT PAST DATA	*NONE**	Name of the set with values of means (and variances) computed previously.
CMVN HYBRID	*false**	Hybrid approach (update means + past vars).
CMVN INIT VAD	*0**	Size (in frames) of the segment around the first non-speech-to-speech transition to be used to estimate initial values of means (and variances).

A.10 VAD

Ahots-aktibitatea detektatzea (VAD) konfiguratzeko parametroak.

Table A.10: VADarekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	Balioa	Deskripzioa
VAD		
VAD	*NONE**	No VAD is used.
	EXTR	VAD based in energy.
	HMM SIL	VAD based on the central state GMM of the silence HMM.
MIN SPEECH FRAMES	*15**	Number of frames for a speech segment to be considered as speech.
MIN SIL FRAMES	*15**	Number of frames for a silence segment to be considered as silence.
SPEECH MARGIN	*0**	Number of silence frames added to a speech segment at both ends.

A.11 UV

Fonemen puntuatzea (UV, *Utterance Verification*) konfiguratzeko parametroak.

Table A.11: UVarekin lotutako parametro konfiguragarriak *AhoSR*n

Parametroa	**Balioa**	**Deskripzioa**
UV		
UV	*false**	Utterance Verification is applied over the Viterbi decoder paths.
UV UNIT	*WORD*	Word-level verification scores are computed.
	*PHONE**	Both word-level and phone-level verification scores are computed.

APPENDIX B

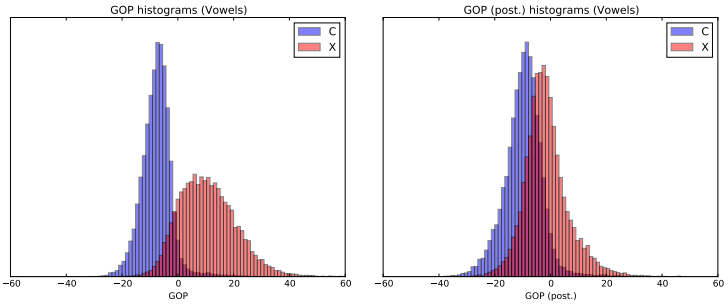
GOPen, iraupenen eta log-egiantzen histogramak

B.1 Zuzen eta oker ebakitako fonemen histogramak

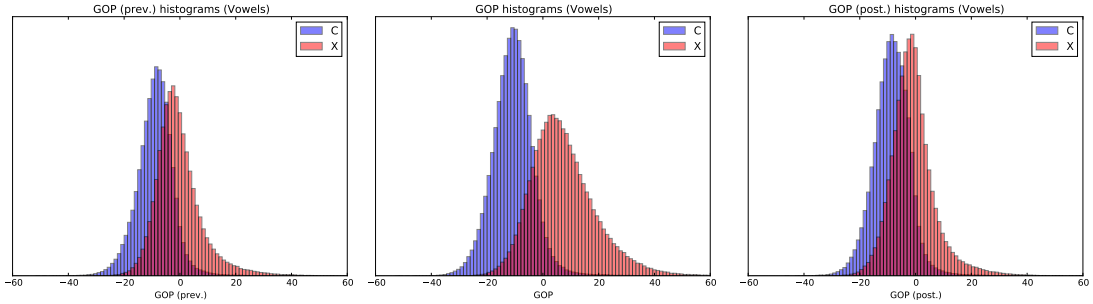
Atal honetan, zuzen (C) eta oker (X) ebakitako fonemen GOPen, iraupenen eta log-egiantzen histogramak jaso ditugu (aurreko fonemarenak eta ondorengoarenak ere bai), fonema-taldearen, iraupenaren eta ahots-segmentuan duen kokapenaren arabera sailkatuta.

1. **BOKALAK**

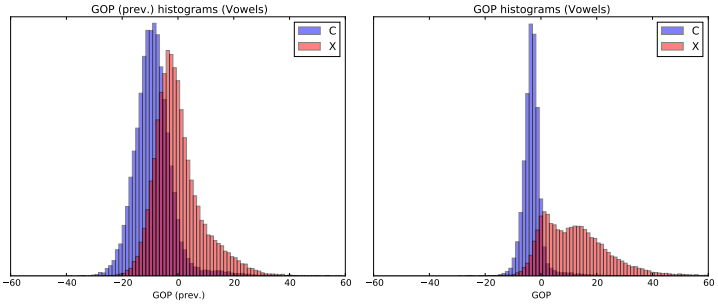
- GOPa:
  - Ezkerreko fonema



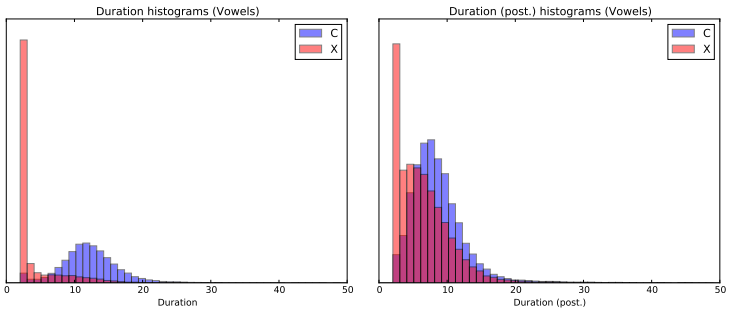
  - Tarteko fonema



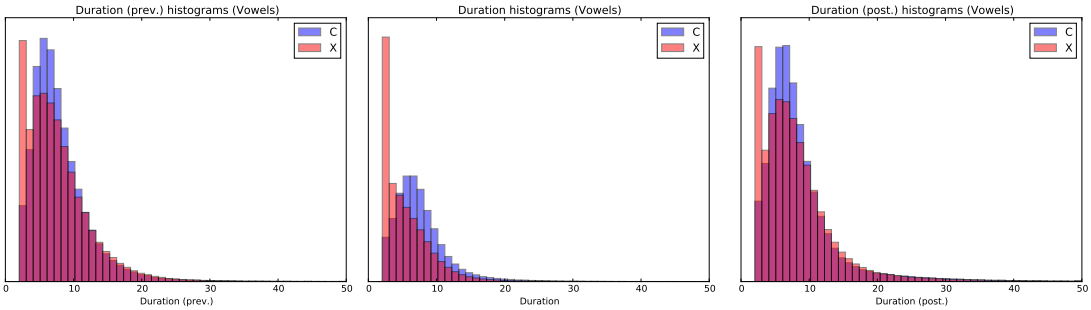
  - Eskuineko fonema



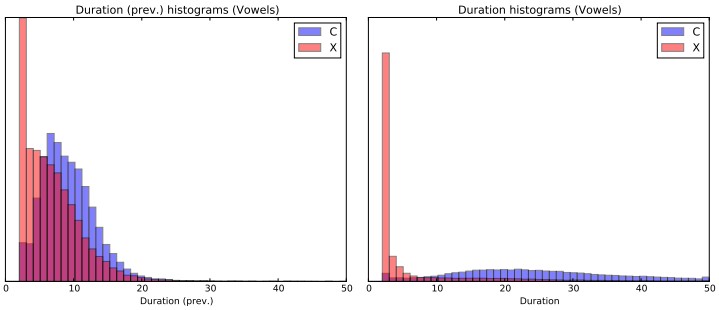
- Iraupena:
  - Ezkerreko fonema



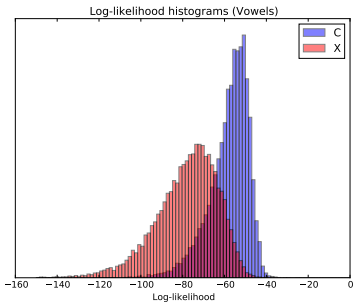
  - Tarteko fonema

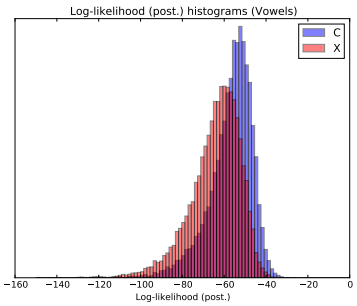


  - Eskuineko fonema

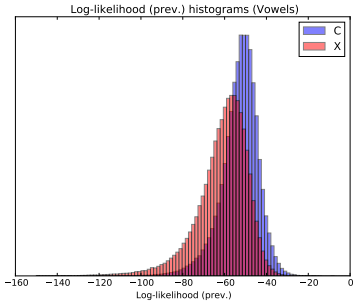


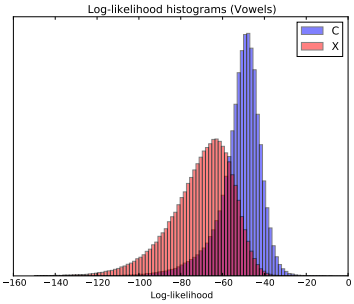
- Log-egiantza:
  - Ezkerreko fonema

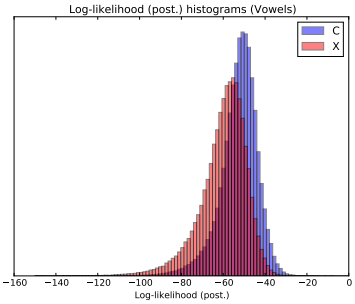




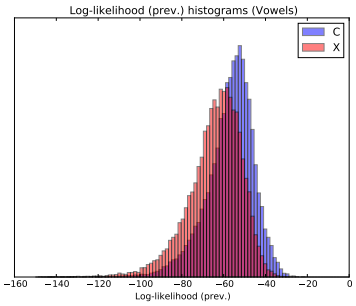
  - Tarteko fonema

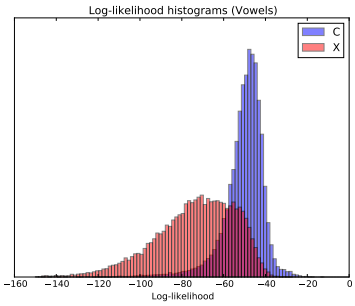






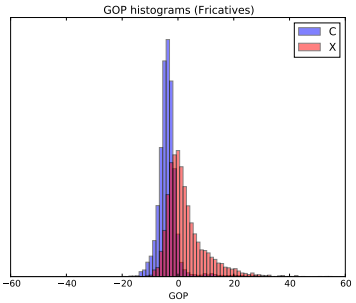
  - Eskuineko fonema

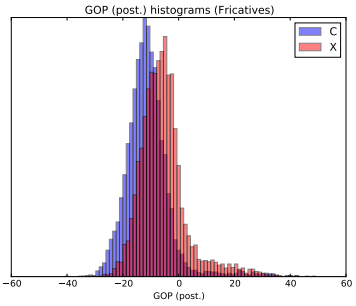




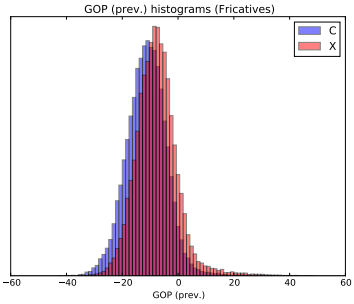
2. IGURZKARIAK

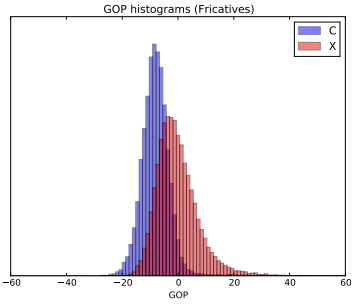
- GOPa:
  - Ezkerreko fonema

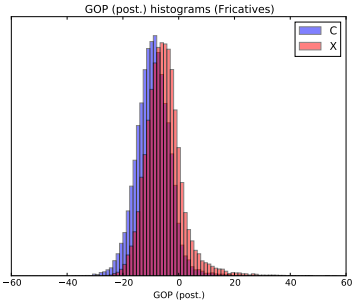




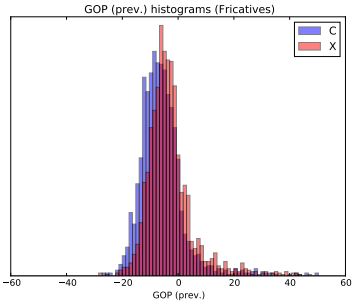
  - Tarteko fonema

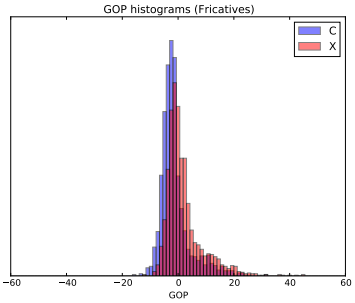




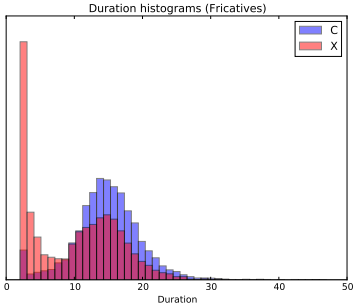


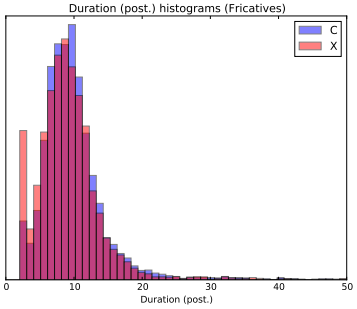
  - Eskuineko fonema



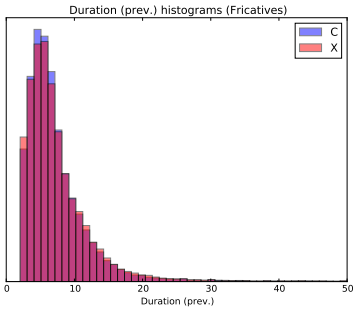


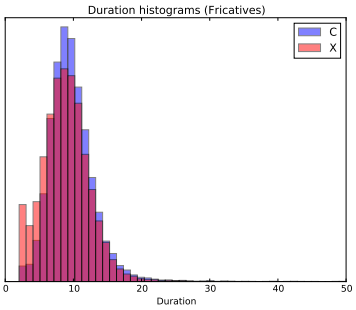
- Iraupena:
  - Ezkerreko fonema

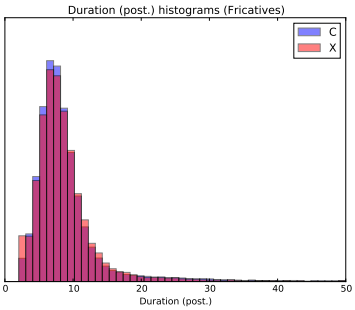




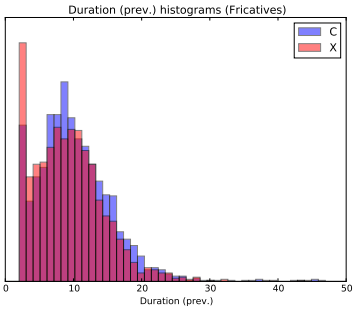
  - Tarteko fonema

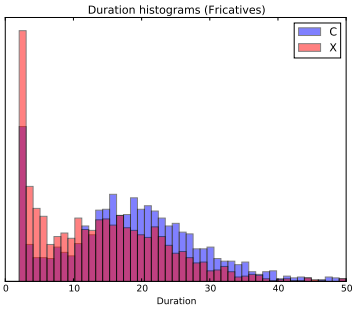




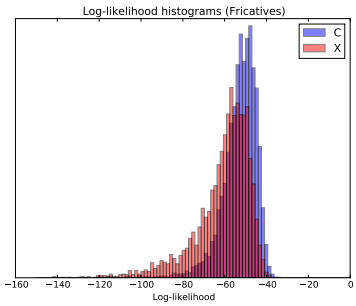


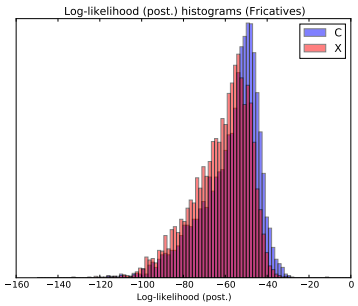
  - Eskuineko fonema



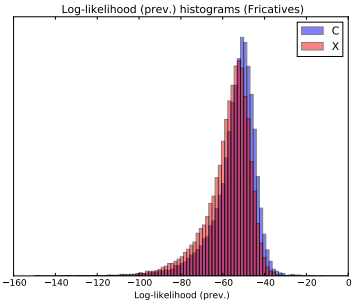


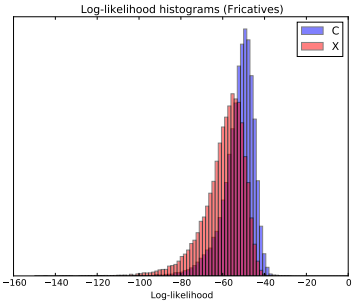
- Log-egiantza:
  - Ezkerreko fonema

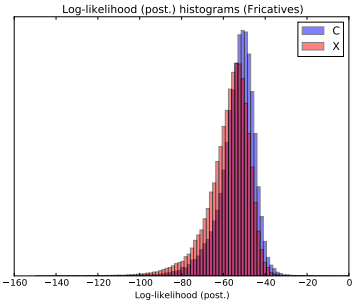




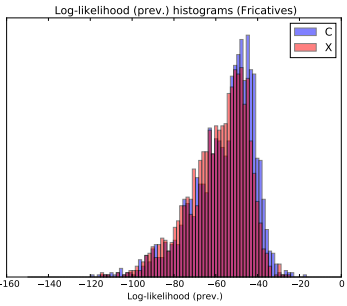
  - Tarteko fonema

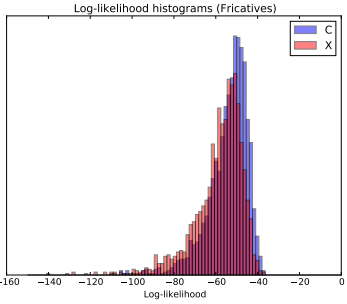






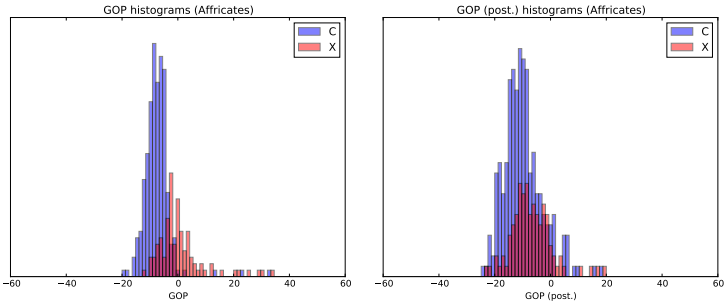
  - Eskuineko fonema



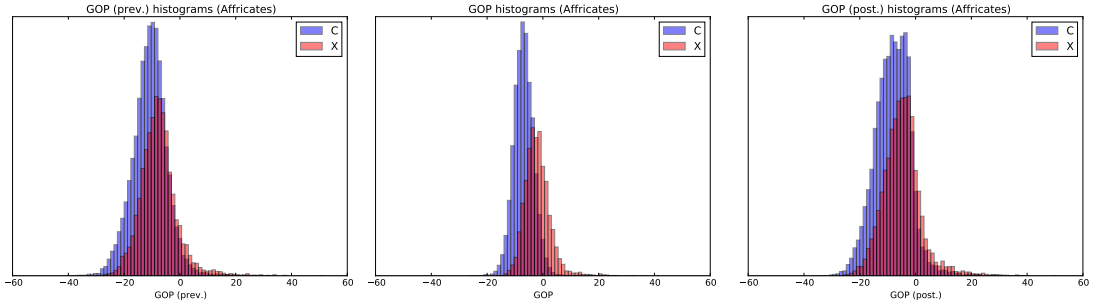


3. AFRIKATUAK

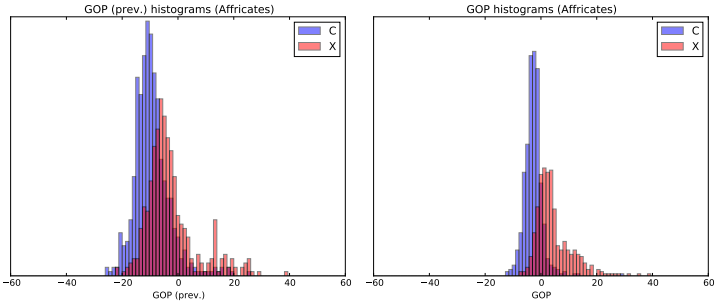
- GOPa:
  - Ezkerreko fonema



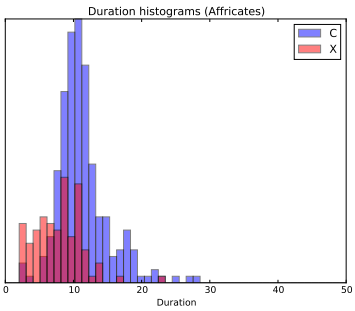
  - Tarteko fonema

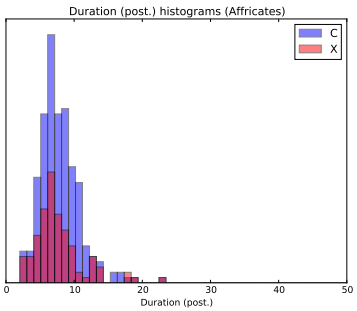


  - Eskuineko fonema

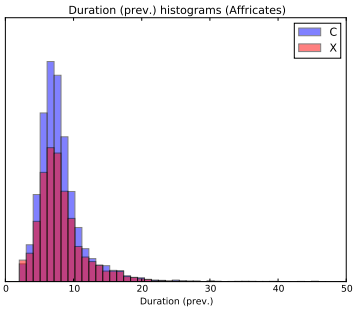


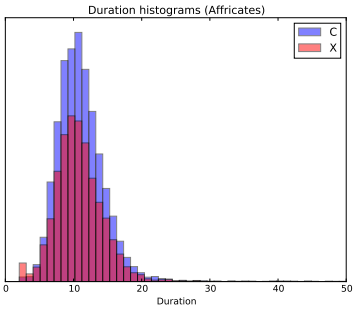
- Iraupena:
  - Ezkerreko fonema

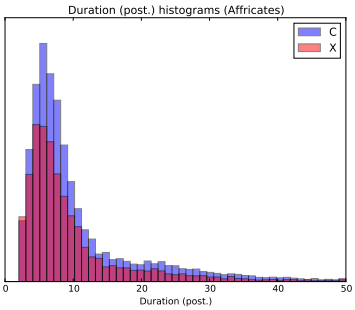




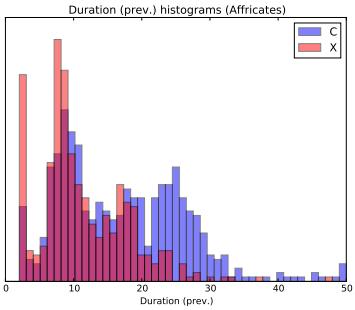
  - Tarteko fonema

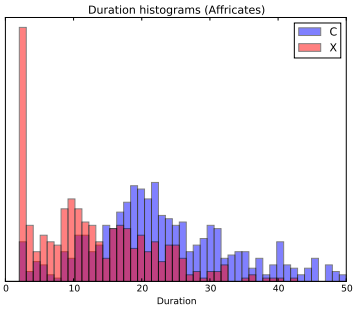




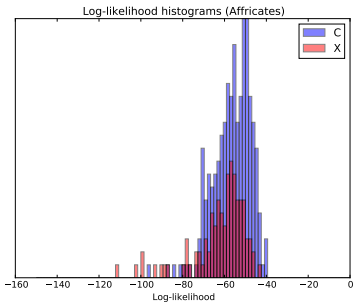


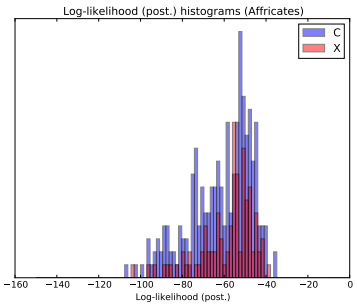
  - Eskuineko fonema



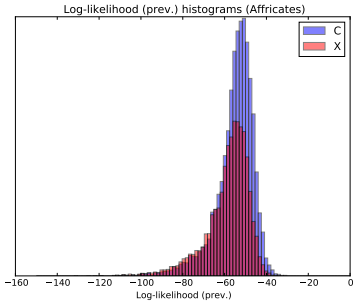


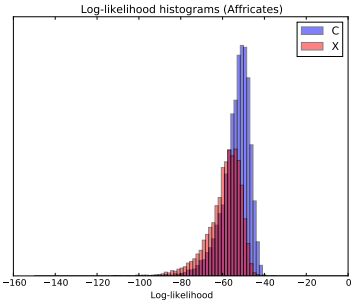
- Log-egiantza:
  - Ezkerreko fonema

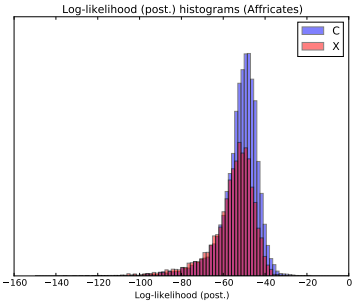




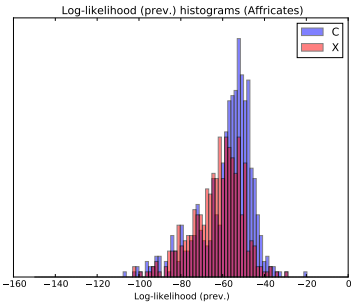
  - Tarteko fonema

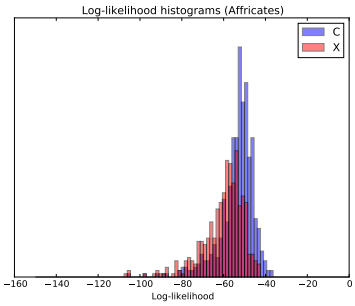






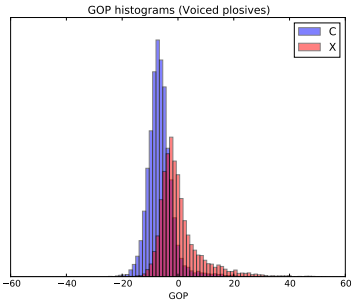
  - Eskuineko fonema

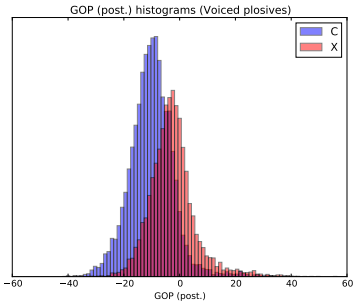




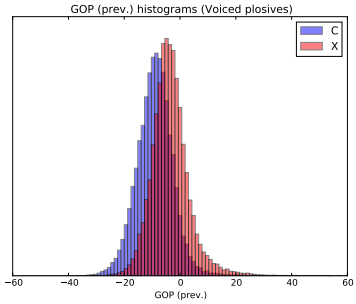
4. HERSKARI AHOSTUNAK

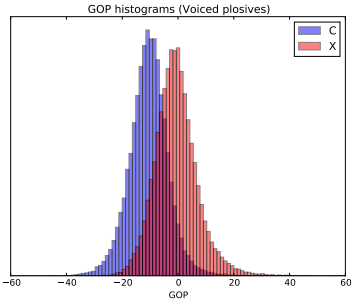
- GOPa:
  - Ezkerreko fonema

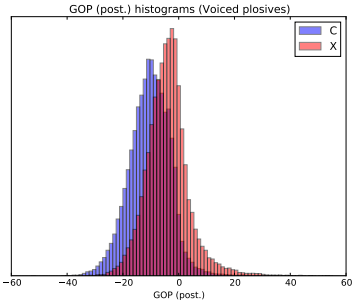




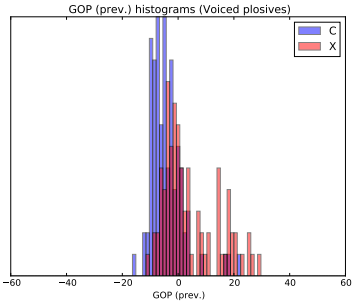
  - Tarteko fonema

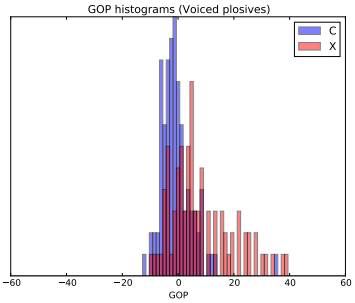




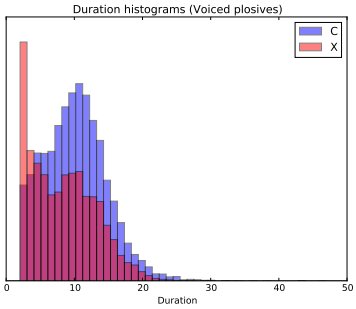


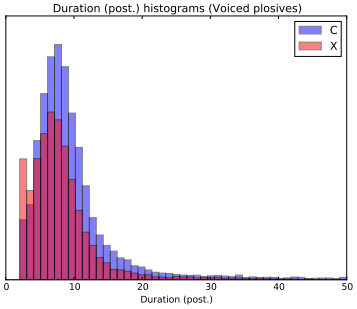
  - Eskuineko fonema



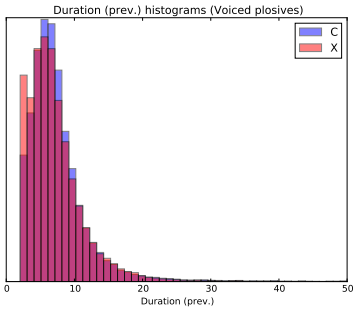


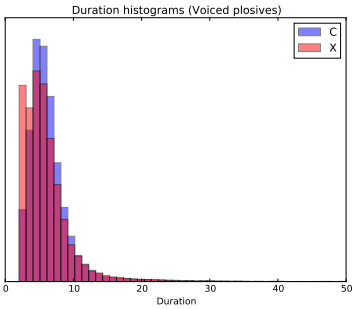
- Iraupena:
  - Ezkerreko fonema

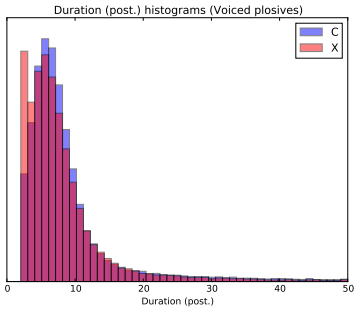




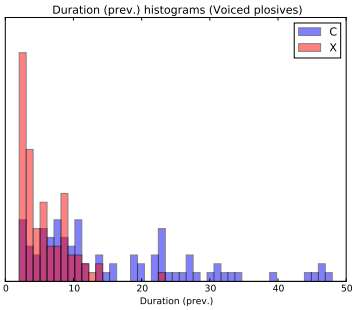
  - Tarteko fonema

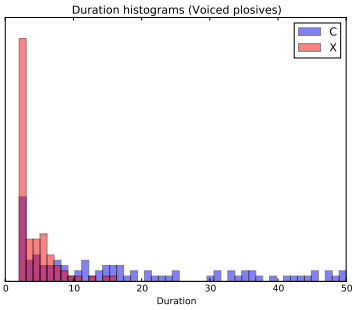




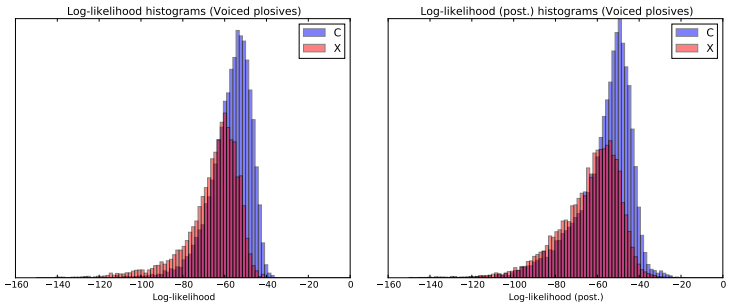


  - Eskuineko fonema

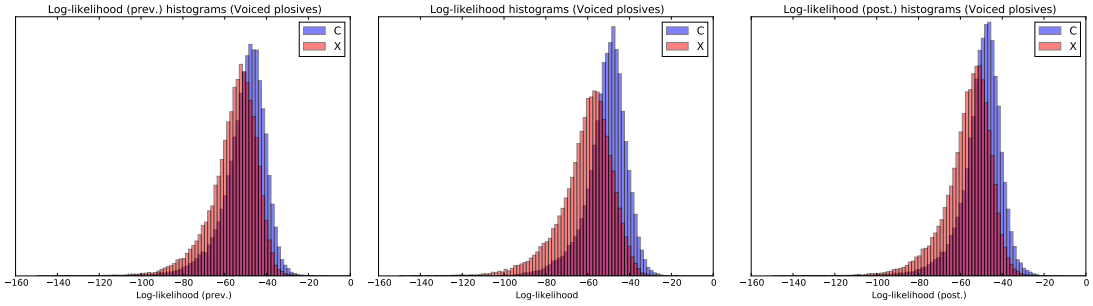




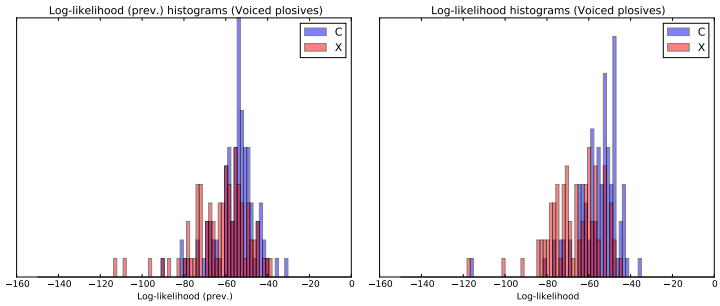
- Log-egiantza:
  - Ezkerreko fonema



  - Tarteko fonema

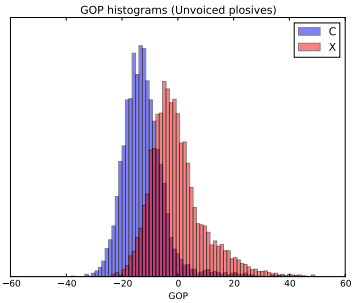


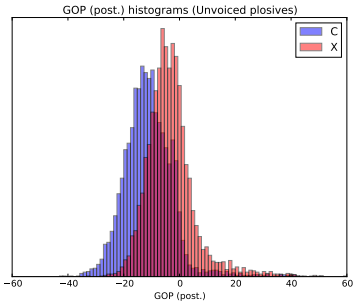
  - Eskuineko fonema



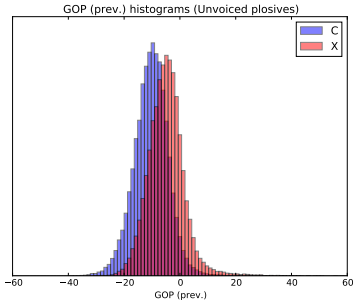
5. HERSKARI AHOSKABEAK

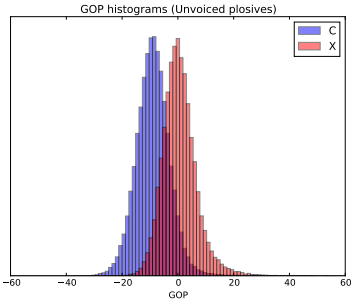
- GOPa:
  - Ezkerreko fonema

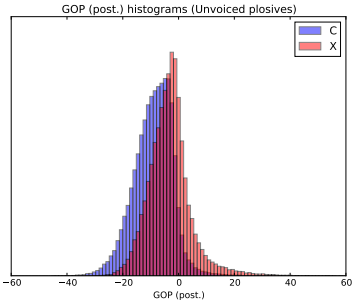




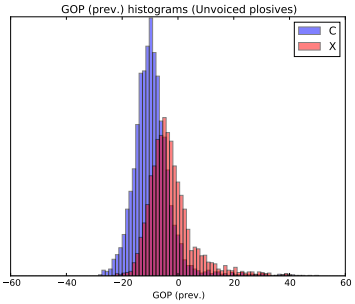
  - Tarteko fonema

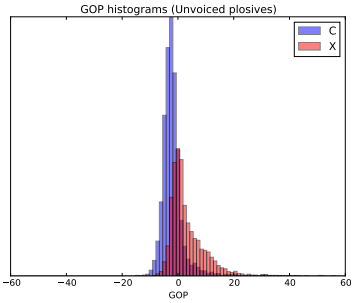




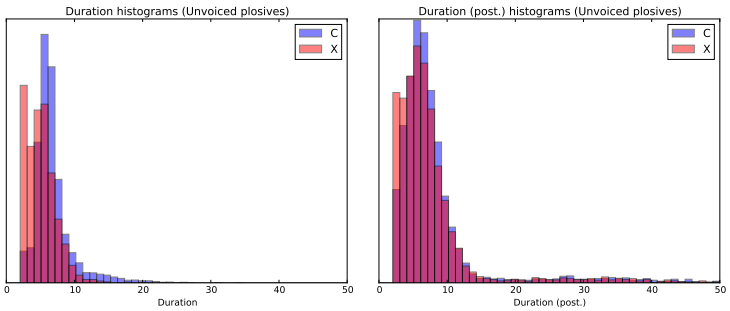


  - Eskuineko fonema

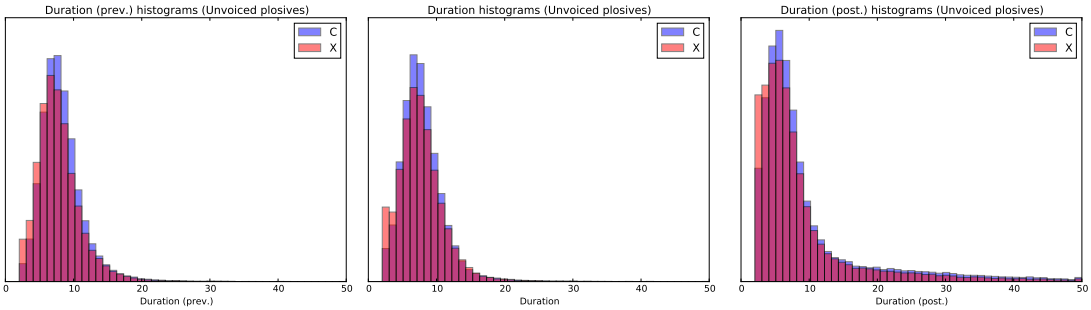




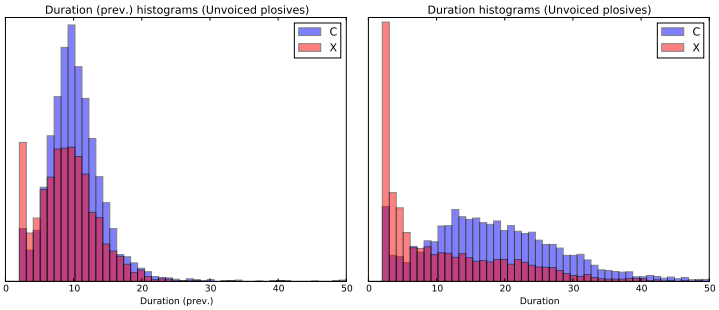
- Iraupena:
  - Ezkerreko fonema



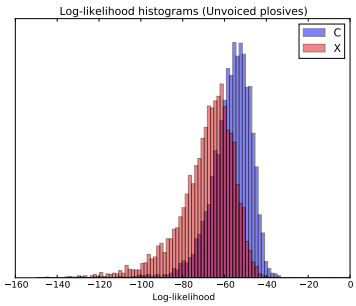
  - Tarteko fonema

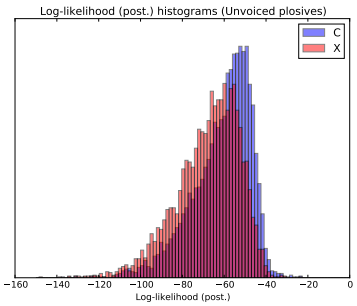


  - Eskuineko fonema

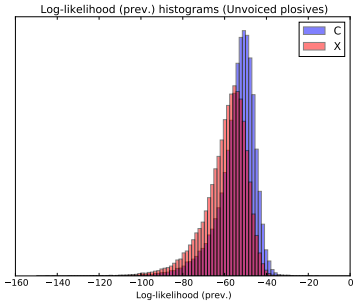


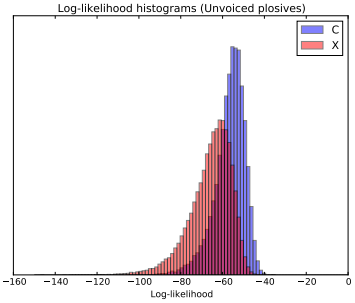
- Log-egiantza:
  - Ezkerreko fonema

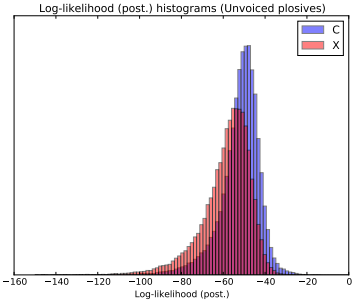




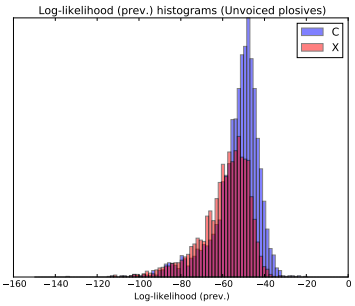
  - Tarteko fonema

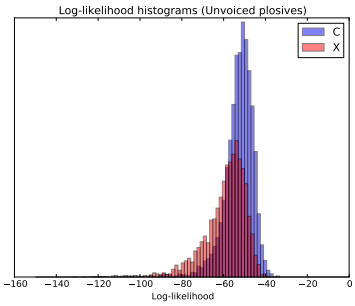






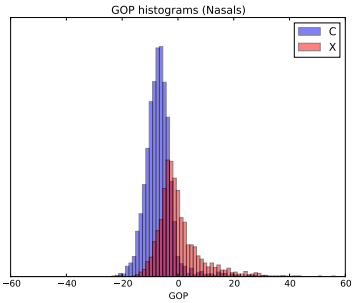
  - Eskuineko fonema

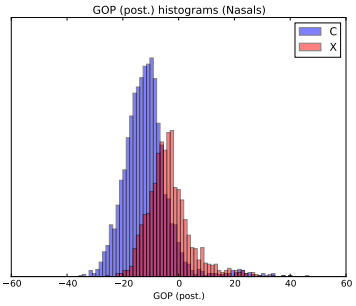




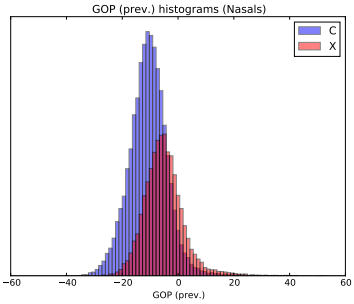
6. **SUDURKARIAK**

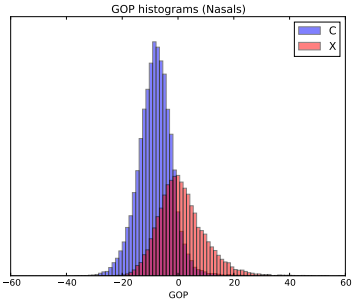
- GOPa:
  - Ezkerreko fonema

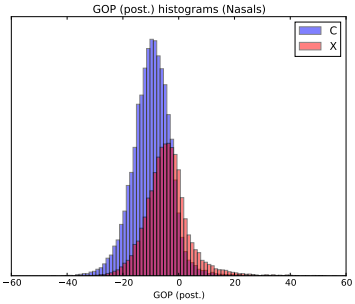




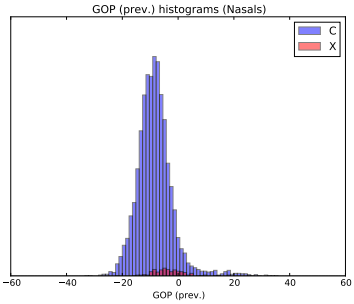
  - Tarteko fonema

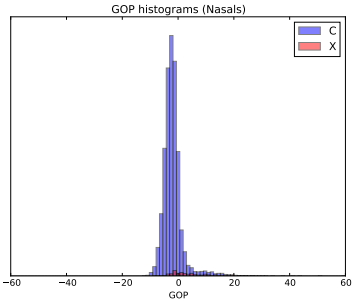




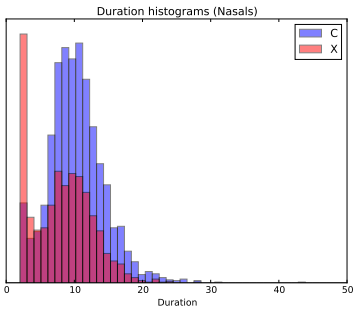


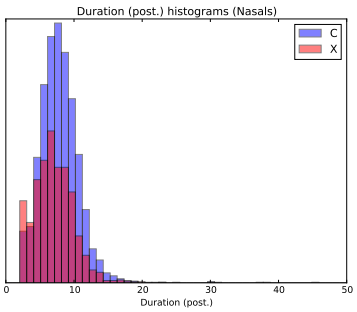
  - Eskuineko fonema



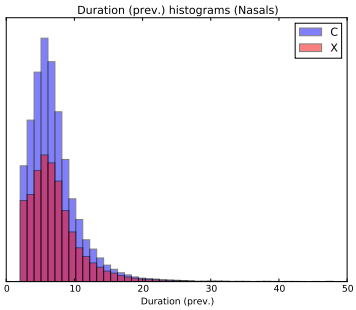


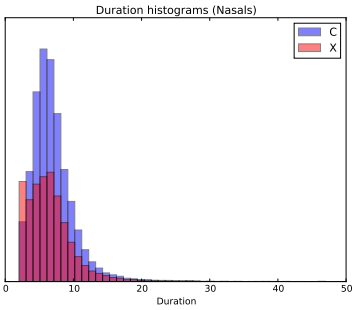
- Iraupena:
  - Ezkerreko fonema

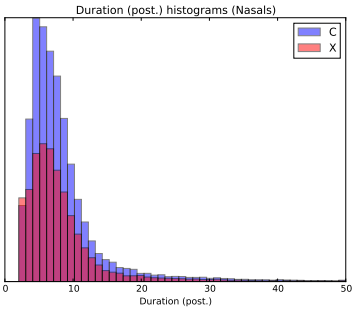




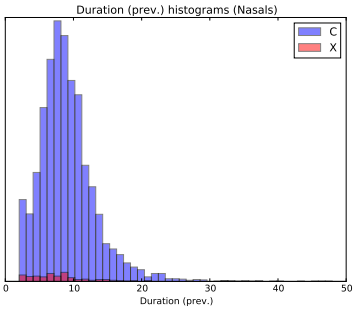
  - Tarteko fonema

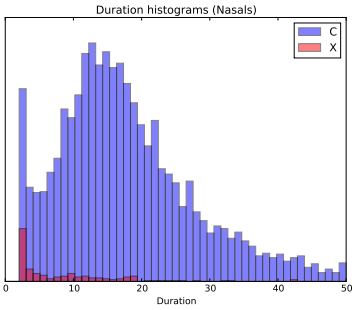




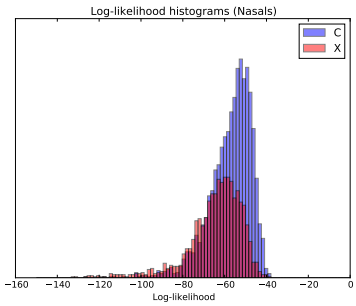


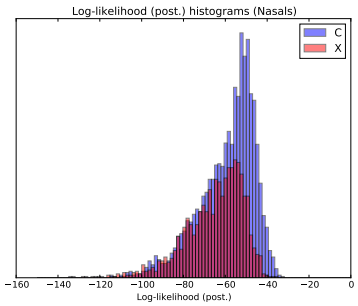
  - Eskuineko fonema



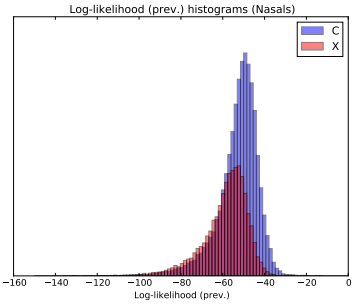


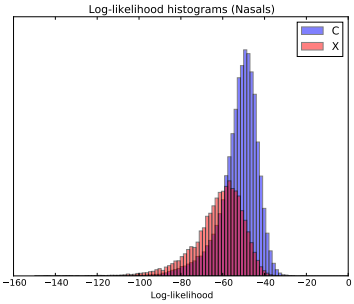
- Log-egiantza:
  - Ezkerreko fonema

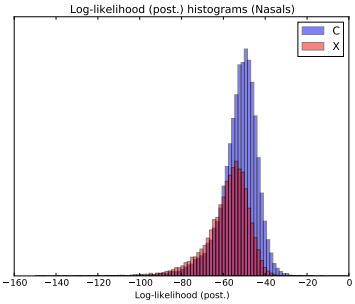




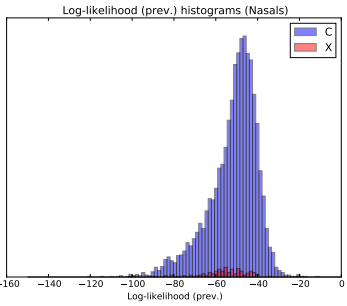
  - Tarteko fonema

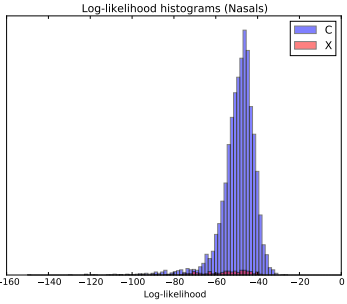






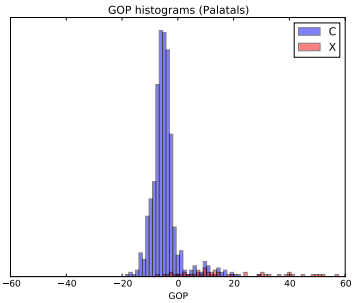
  - Eskuineko fonema

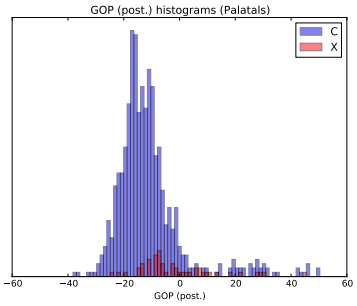




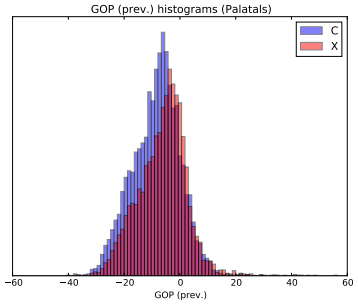
7. SABAIKARIAK

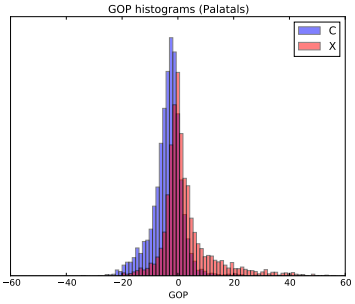
- GOPa:
  - Ezkerreko fonema

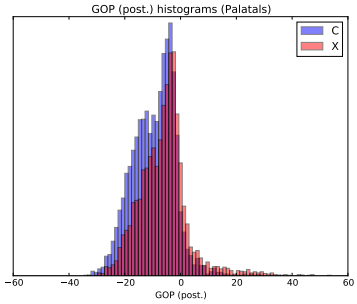




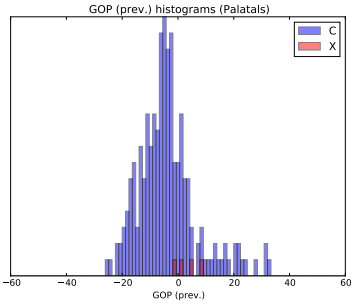
  - Tarteko fonema

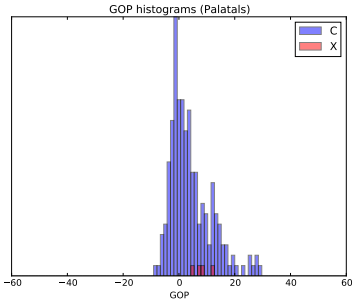




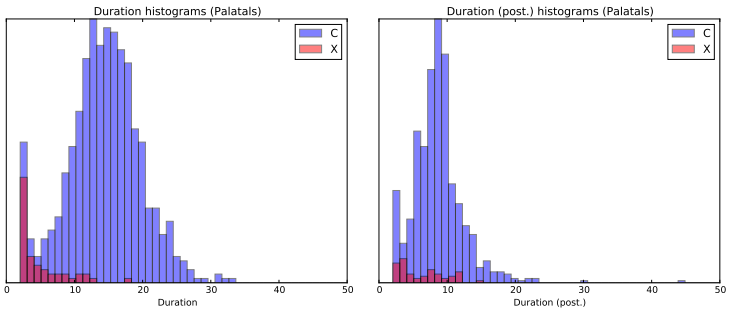


  - Eskuineko fonema

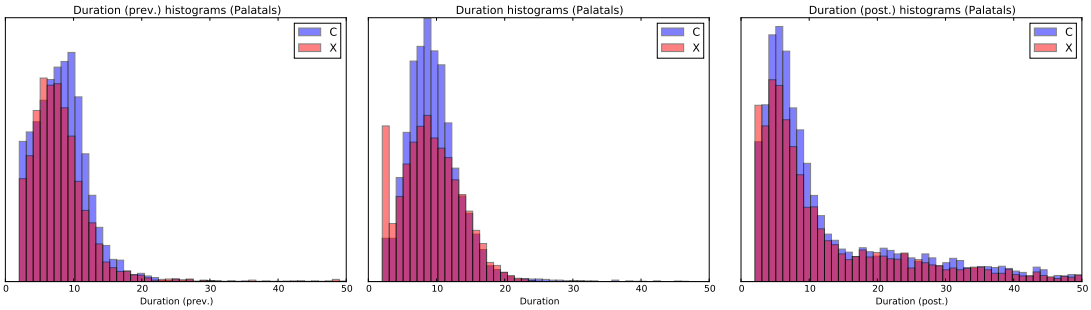




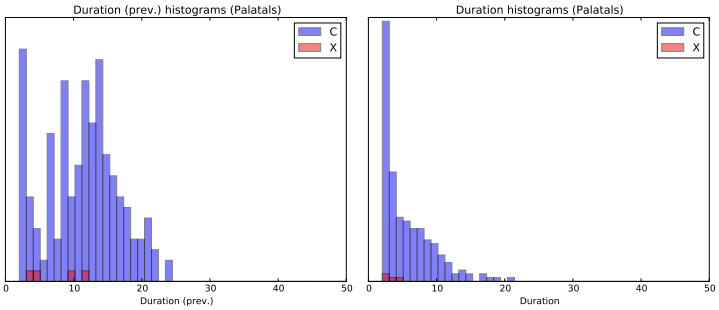
- Iraupena:
  - Ezkerreko fonema



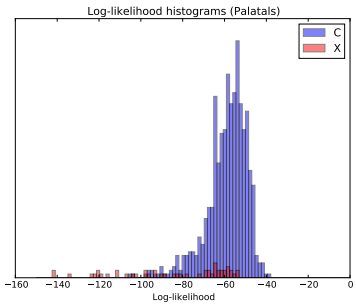
  - Tarteko fonema

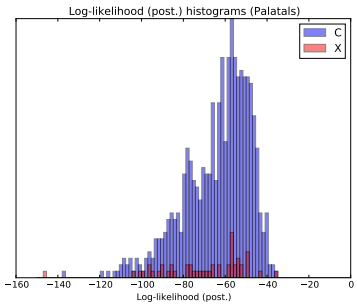


  - Eskuineko fonema

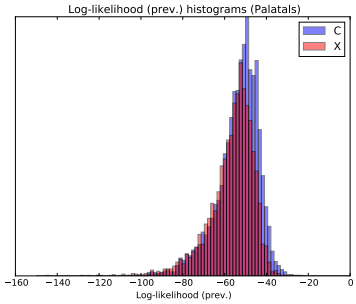


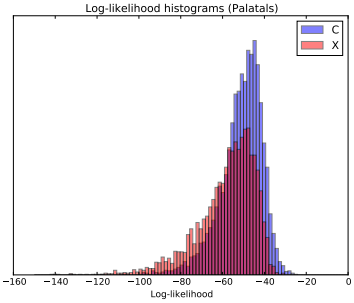
- Log-egiantza:
  - Ezkerreko fonema

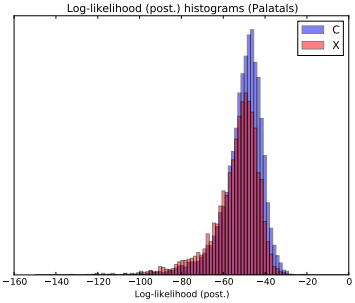




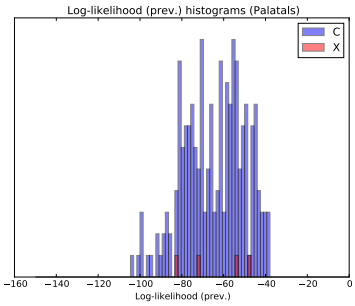
  - Tarteko fonema

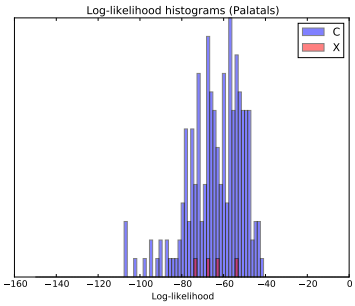






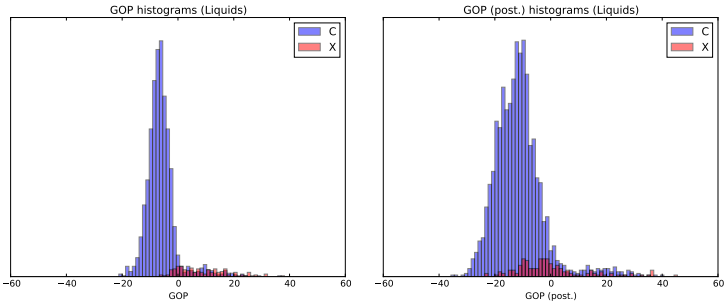
  - Eskuineko fonema



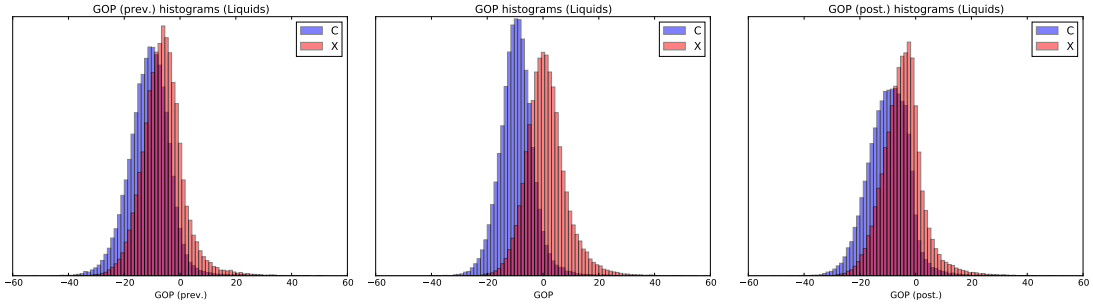


8. **LIKIDOAK**

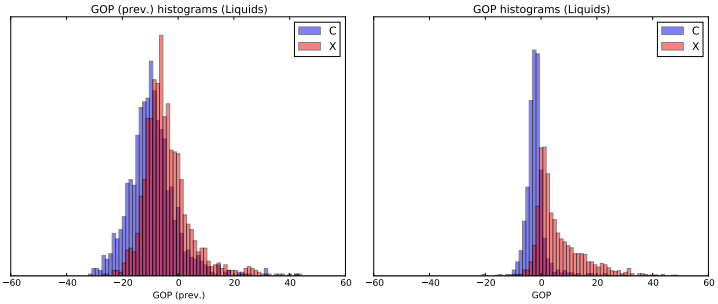
- GOPa:
  - Ezkerreko fonema



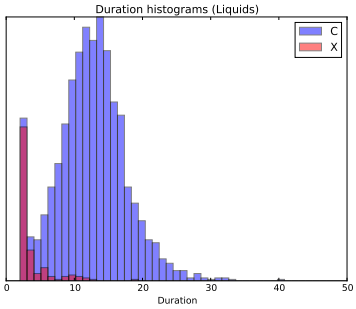
  - Tarteko fonema

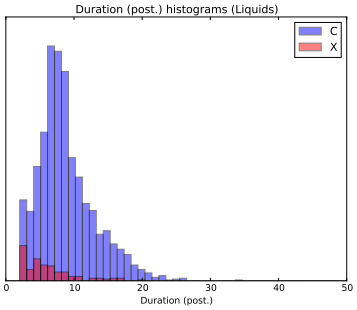


  - Eskuineko fonema

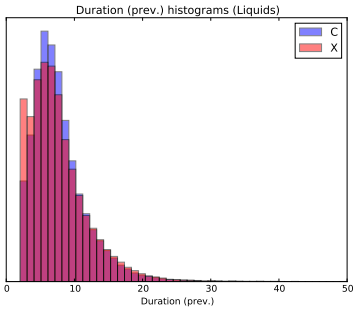


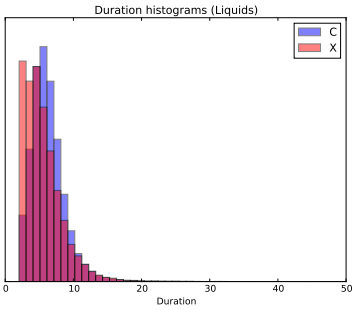
- Iraupena:
  - Ezkerreko fonema

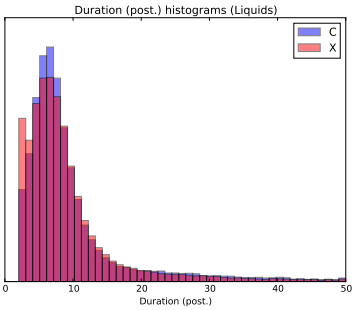




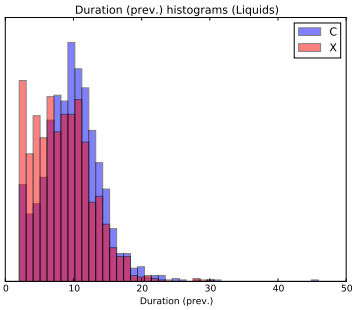
  - Tarteko fonema

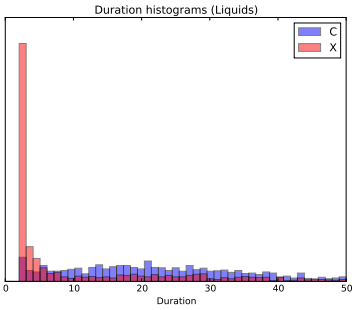




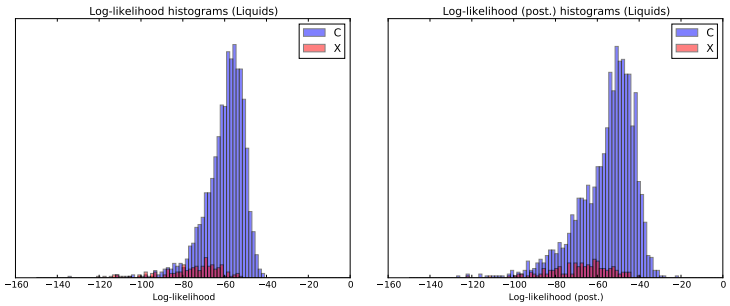


  - Eskuineko fonema

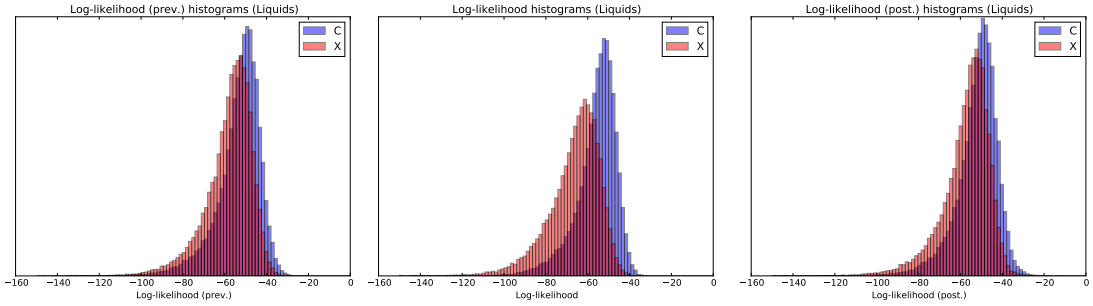




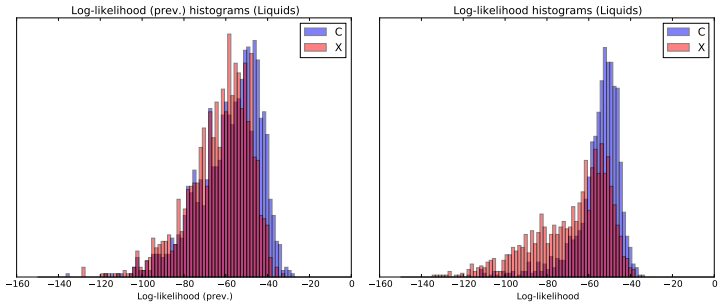
- Log-egiantza:
  - Ezkerreko fonema



  - Tarteko fonema



  - Eskuineko fonema



Bibliography

[1] P. Robinson, *The Routledge Encyclopedia of Second Language Acquisition*. Routledge, 2012.

[2] M. Levy and G. Stockwell, *CALL dimensions: options and issues in Computer-Assisted Language Learning*. Taylor & Francis, 2nd ed., 2013.

[3] B. G. Vice Ministry for Language Policy, "Fifth sociolinguistic survey," 2013.

[4] J. I. Hualde and K. Zuazo, "The standardization of the basque language," *Language Problems and Language Planning*, vol. 31, no. 2, pp. 143–168, 2007.

[5] A. Alberdi and E. A. Batzordea, *Ahoskera*. Hizkuntza Prestakuntza, Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia, 2014.

[6] J. Hualde, *Basque Phonology*. Taylor & Francis, 2004.

[7] M. E. Butler-Pascoe, "The history of call: The intertwining paths of technology and secondforeign language teaching," *International Journal of Computer-Assisted Language Learning and Teaching*, vol. 1, no. 1, pp. 16–32, 2011.

[8] J. Harmer, "The practice of english language teaching," *London/New York*, 1991.

[9] C. A. Chapelle, "English language learning and technology," *Language Learning & Language Teaching*, vol. 7, 2003.

[10] C. Lai, "Modeling teachers' influence on learners' self-directed use of technology for language learning outside the classroom," *Computers & Education*, vol. 82, pp. 74–83, 2015.

[11] P. Benson, *Teaching and researching: Autonomy in language learning*. Routledge, 2013.

[12] C. Dorothy, K. Richard, and S. Bryan, "Technology in language use, language teaching, and language learning," *The Modern Language Journal*, vol. 100, no. S1, pp. 64–80, 2016.

[13] R. Kern, P. Ware, and M. Warschauer, *Encyclopedia of Language and Education*, ch. Network-Based Language Teaching, pp. 1374–1385. Springer US, 2008.

[14] S. Herring, D. Stein, and T. Virtanen, *Pragmatics of Computer-Mediated Communication*. Handbooks of Pragmatics (HOPS), De Gruyter, 2013.

[15] S. Fotos and C. Browne, *New Perspectives on CALL for Second Language Classrooms*, ch. Teaching WELL and Loving It. Taylor & Francis, 2013.

[16] O. Viberg and k. Grönlund, "Mobile assisted language learning: A literature review," in *mLearn*, vol. 955, pp. 9–16, 2012.

[17] M. Thomas, H. Reinders, and M. Warschauer, *Contemporary Computer-Assisted Language Learning*, ch. Intelligent CALL. Bloomsbury linguistics, Bloomsbury Academic, 2012.

[18] A. Davies, "Computer-assisted language testing," *CALICO Journal*, vol. 1, no. 5, pp. 41–43, 2013.

[19] D. R. Garrison, *E-learning in the 21st century: A framework for research and practice*. Taylor & Francis, 2011.

[20] N. Nagata, "Computer vs. workbook instruction in second language acquisition," *CALICO Journal*, vol. 14, no. 1, 1996.

[21] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[22] J. S. Payne and P. Whitney, "Developing l2 oral proficiency through synchronous cmc: Output, working memory, and interlanguage development," *CALICO journal*, vol. 20, pp. 7–32, 01 2002.

[23] U. Felix, "The unreasonable effectivness of call: What have we learned in two decades of research?," *ReCALL*, vol. 20, no. 2, p. 141–161, 2008.

[24] R. Clifford and N. Granoien, *The path of speech technologies in Computer Assisted Language Learning: From research toward practice*, ch. Applications of technology to language acquisition processes: What can work and why, pp. 25–43. Routledge, 2008.

[25] E. M. Golonka, A. R. Bowles, V. M. Frank, D. L. Richardson, and S. Freynik, "Technologies for foreign language learning: a review of technology types and their effectiveness," *Computer Assisted Language Learning*, vol. 27, no. 1, pp. 70–105, 2014.

[26] S. Bodnar, C. Cucchiarini, and H. Strik, "Computer-assisted grammar practice for oral communication," in *Proc. of International Conference on Computer Supported Education (CSEDU)*, vol. 1, (Noordwijkerhout, Netherlands), pp. 355–361, 2011.

[27] M. Eskénazi, "Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype," *Language Learning & Technology*, vol. 2, no. 2, pp. 67–76, 1999.

[28] W. L. Johnson, S. Marsella, N. Mote, and H. Viljhálmsson, "Tactical language training system: Supporting the rapid acquisition of foreign language and cultural skills," in *Proc. of InSTILL/ICALL Symposium: NLP and speech technologies in advanced language learning systems (InSTIL/ICALL)*, (Venice, Italy), 2004.

[29] O. Jokisch, U. Koloska, D. Hirschfeld, and R. Hoffmann, "Pronunciation learning and foreign accent reduction by an audiovisual feedback system," in *Affective Computing and Intelligent Interaction (ACII)*, (Berlin, Germany), pp. 419–425, Springer, Berlin, Heidelberg, 2005.

[30] H. Morton and M. A. Jack, "Scenario-based spoken interaction with virtual agents," *Computer Assisted Language Learning*, vol. 18, no. 3, pp. 171–191, 2005.

[31] H. Morton and M. Jack, "Speech interactive computer-assisted language learning: a cross-cultural evaluation," *Computer Assisted Language Learning*, vol. 23, no. 4, pp. 295–319, 2010.

[32] M. Eskenazi, A. Kennedy, C. Ketchum, R. Olszewski, and G. Pelton, "The nativeaccenttm pronunciation tutor: measuring success in the real world," in *Proc. of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, (Farmington, USA), pp. 124–127, International Speech Communication Association (ISCA), 2007.

[33] S. Chevalier, "Speech interaction with saybot, a call software to help chinese learners of english," in *Proc. of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, (Farmington, USA), pp. 37–40, International Speech Communication Association (ISCA), 2007.

[34] D. Grazyna, A. Wagner, N. Cylwik, and O. Jokisch, "An audiovisual feedback system for acquiring l2 pronunciation and l2 prosody," in *Proc. of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, (Warwickshire, UK), International Speech Communication Association (ISCA), 2009.

[35] H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda, "Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.

[36] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *Proc. of International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, (Stockholm, Sweden), pp. 1–8, International Speech Communication Association (ISCA), 2012.

[37] H. Hamada, S. Miki, and R. Nakatsu, "Automatic evaluation of english pronunciation based on speech recognition techniques," *IEICE Transactions on Information and Systems*, vol. E76-D, no. 3, pp. 352–359, 1993.

[38] S. M. Hiller, E. Rooney, J. Laver, and M. A. Jack, "Spell: An automated system for computer-aided pronunciation teaching," *Speech Communication*, vol. 13, no. 3–4, pp. 463–473, 1993.

[39] S. M. A. Goddijn and G. de Krom, "Evaluation of second language learners' pronunciation using hidden Markov models," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Rhodes, Greece), pp. 2331–2334, 1997.

[40] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic evaluation and training in English pronunication," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, (Kobe, Japan), pp. 1185–1188, 1990.

[41] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Pronunciation scoring of foreign language student speech," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, (Philadelphia, USA), pp. 1457–1460, 1996.

[42] G. Kawai and K. Hirose, "A CALL system using speech recognition to train the pronunciation of japanese long vowels, the mora nasal and mora obstruent," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Rhodes, Greece), pp. 657–660, 1997.

[43] O. Ronen, L. Neumeyer, and H. Franco, "Automatic detection of mispronunciation for language instruction," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Rhodes, Greece), pp. 649–652, International Speech Communication Association (ISCA), 1997.

[44] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Rhodes, Greece), pp. 645–648, International Speech Communication Association (ISCA), 1997.

[45] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Budapest, Hungary), pp. 851–854, International Speech Communication Association (ISCA), 1999.

[46] S. M. Witt, *Use of speech recognition in Computer-assisted Language Learning.* PhD dissertation, University of Cambridge, Department of Engineering, 1999.

[47] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for mandarin," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Las Vegas, USA), pp. 5077–5080, Institute of Electrical and Electronics Engineers (IEEE), 2008.

[48] Y. Wang and L. lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Kyoto, Japan), pp. 5049–5052, Institute of Electrical and Electronics Engineers (IEEE), 2012.

[49] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 11, (Tokyo, Japan), pp. 49–52, Institute of Electrical and Electronics Engineers (IEEE), 1986.

[50] B.-H. Juang, W. Hou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[51] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, (Orlando, USA), pp. I–105–I–108, Institute of Electrical and Electronics Engineers (IEEE), 2002.

[52] K. Yan and S. Gong, "Pronunciation proficiency evaluation based on discriminatively refined acoustic models," *International Journal of Information Technology and Computer Science*, vol. 3, no. 2, pp. 17–23, 2011.

[53] X. Qian, F. Soong, and H. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (capt)," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Makuhari, Japan), pp. 757–760, International Speech Communication Association (ISCA), 2010.

[54] A. Ito, Y.-L. Lim, M. Suzuki, and S. Makino, "Pronunciation error detection method based on error rule clustering using a decision tree," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Lisbon, Portugal), pp. 173–176, International Speech Communication Association (ISCA), 2005.

[55] K. Truong, *Automatic Pronunciation Error Detection in Dutch as a Second Language: An Acoustic–phonetic Approach*. PhD dissertation, Utrecht University, Faculty of Humanities, 2004.

[56] H. Strik, K. Truong, F. de Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[57] H. Jiang, "Confidence measures for speech recognition: A surve," *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[58] S. Wei, G. Hu, Y. Hu, and R.-H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.

[59] J. Jiang and B. Xu, "Exploring the automatic mispronunciation detection of confusable phones for mandarin," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Taipei, Taiwan), pp. 4833–4836, Institute of Electrical and Electronics Engineers (IEEE), 2009.

[60] S. Xu, J. Jiang, Z. Chen, and B. Xu, "Automatic pronunciation error detection based on linguistic knowledge and pronunciation space," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Taipei, Taiwan), pp. 4841–4844, Institute of Electrical and Electronics Engineers (IEEE), 2009.

[61] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Landmark-based automated pronunciation error detection," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Makuhari, Japan), pp. 614–617, International Speech Communication Association (ISCA), 2010.

[62] K. Hirabayashi and S. Nakagawa, "Automatic evaluation of english pronunciation by japanese speakers using various acoustic features and pattern recognition techniques," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Makuhari, Japan), pp. 598–601, International Speech Communication Association (ISCA), 2010.

[63] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[64] W. Hu, Y. Qian, and F. Soong, "A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call)," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Lyon,

France), pp. 1886–1890, International Speech Communication Association (ISCA), 2013.

[65] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[66] X. Qian, H. Meng, and F. K. Soong, "The use of dbn-hmms for mispronunciation detection and diagnosis in l2 english to support computer-aided pronunciation training," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Portland, USA), pp. 775–778, International Speech Communication Association (ISCA), 2012.

[67] I. Odriozola, O. Jokisch, I. Hernaez, and R. Hoffmann, "A pronunciation tutoring system for basque - first development steps," in *Proc. of ESSV (Elektronische Sprachsignalverarbeitung)*, (Cottbus, Germany), 2012.

[68] J. Bernstein, J. Cheng, and M. Suzuki, "Fluency changes with general progress in l2 proficiency," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Makuhari, Japan), International Speech Communication Association (ISCA), 2010.

[69] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: an automatic approach," *The Journal of the Acoustical Society of America*, vol. 107, pp. 989–999, 1998.

[70] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of failure: employing word accent information for pronunciation quality assessment of english l2 learners," in *Proc. of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, (Warwickshire, UK), International Speech Communication Association (ISCA), 2009.

[71] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody – annotation, modelling and evaluation," in *Proc. of International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, (Stockholm, Sweden), pp. 21–30, International Speech Communication Association (ISCA), 2012.

[72] A. Bonneau and V. Colotte, "Automatic feedback for l2 prosody learning," in *Speech and Language Technologies* (I. Ipsic, ed.), pp. 55–70, Intech, 2011.

[73] G.-A. Levow, "Investigating pitch accent recognition in non-native speech," in *Proc. of the ACL-IJCNLP 2009 Conference Short Papers*, (Stroudsburg, USA), pp. 269–272, Association for Computational Linguistics, 2009.

[74] A. Ito, Y.-L. Lim, M. Suzuki, and S. Makino, "Pronunciation error detection for computer-assisted language learning system based on error rule clustering using a decision tree," *Acoustical Science and Technology*, vol. 28, no. 2, pp. 131–133, 2007.

[75] H. Ye and S. Young, "Improving speech recognition performance of beginners in spoken conversational interaction for language learning," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Lisbon, Portugal), International Speech Communication Association (ISCA), 2005.

[76] O. Saz, E. Lleida, and W. R. Rodríguez, "Acoustic-phonetic decoding for assessment of mispronunciations in speakers with cognitive disorders," in *Proc. of Advanced Voice Function Assessment (AVFA) International Workshop*, (Madrid, Spain), 2009.

[77] O. Husby, s. Øvregaard, P. Wik, y. Bech, E. Albertsen, S. Nefzaoui, E. Skarpnes, and J. Koreman, "Dealing with l1 background and l2 dialects in norwegian capt," in *Proc. of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, (Venice, Italy), International Speech Communication Association (ISCA), 2011.

[78] A. Neri, C. Cucchiarini, and H. Strik, "Segmental errors in dutch as a second language: how to establish priorities for capt," in *Proc. of InSTILL/ICALL Symposium: NLP and speech technologies in advanced language learning systems (InSTIL/ICALL)*, (Venice, Italy), 2004.

[79] W. K. Lo, S. Zhang, and H. M. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Makuhari, Japan), International Speech Communication Association (ISCA), 2010.

[80] A. M. Harrison, W. Y. Lau, H. M. Meng, and L. Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Brisbane, Australia), pp. 2787–2790, International Speech Communication Association (ISCA), 2008.

[81] A. M. Harrison, W.-k. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *Proc. of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, (Warwickshire, UK), International Speech Communication Association (ISCA), 2009.

[82] X. Qian, H. Meng, and F. Soong, "On mispronunciation lexicon generation using joint-sequence multigrams in computer-aided pronunciation training (capt)," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Makuhari, Japan), International Speech Communication Association (ISCA), 2010.

[83] J. M. Norris and L. Ortega, "Effectiveness of l2 instruction: A research synthesis and quantitative meta-analysis," *Language learning*, vol. 50, no. 3, pp. 417–528, 2000.

[84] N. C. Ellis and P. S. Bogart, "Speech and language technology in education: the perspective from sla research and practice.," in *Proc. of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, (Farmington, USA), pp. 1–8, International Speech Communication Association (ISCA), 2007.

[85] F. Ehsani, J. Bernstein, A. Najmi, and O. Todic, "Subarashii: Japanese interactive spoken language education," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Rhodes, Greece), pp. 681–684, International Speech Communication Association (ISCA), 1997.

[86] J. Bernstein, A. Najmi, and F. Ehsani, "Subarashii: Encounters in japanese spoken language education," *CALICO journal*, vol. 16, no. 3, pp. 361–384, 1999.

[87] A. Raux and M. Eskenazi, "Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges," in *Proc. of InSTILL/ICALL Symposium: NLP and speech technologies in advanced language learning systems (InSTIL/ICALL)*, (Venice, Italy), 2004.

[88] J. Lee and S. Seneff, "Automatic grammar correction for second-language learners," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Pittsburgh, USA), pp. 1978–1981, International Speech Communication Association (ISCA), 2006.

[89] H. Wang, C. J. Waple, and T. Kawahara, "Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition," *Speech Communication*, vol. 51, no. 10, pp. 995–1005, 2009.

[90] H. Strik, J. van de Loo, J. van Doremalen, and C. Cucchiarini, "Practicing syntax in spoken interaction: Automatic detection of syntactical errors in non-native utterances," in *Proc. of Interspeech Second Language Studies Workshop*, (Tokyo, Japan), 2010.

[91] H. Strik, J. Colpaert, J. van Doremalen, and C. Cucchiarini, "The disco asr-based call system: practicing l2 oral skills and beyond," in *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pp. 2702–2707, European Language Resource Association (ELRA), 2012.

[92] B. Penning de Vries, C. Cucchiarini, S. Bodnar, H. Strik, and R. van Hout, "Spoken grammar practice and feedback in an asr-based call system," *Computer Assisted Language Learning*, vol. 28, no. 6, pp. 550–576, 2014.

[93] J. Van Doremalen, H. Strik, and C. Cucchiarini, "Utterance verification in language learning applications," in *Proc. of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, (Warwickshire, UK), pp. 13–16, International Speech Communication Association (ISCA), 2009.

[94] G. Bouwman and L. Boves, "Utterance verification based on the likelihood distance to alternative paths," in *Text, Speech and Dialogue*, (Berlin, Heidelberg), pp. 213–220, Springer Berlin Heidelberg, 2002.

[95] I. Odriozola, I. Hernaez, and E. Navas, "Design of a message verification tool to be implemented in call systems," in *Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH)*, (Madrid, Spain), pp. 251–259, 2012.

[96] I. Odriozola, I. Hernaez, and E. Navas, "An on-line VAD based on Multi-Normalisation Scoring (MNS) of observation likelihoods," *Expert Systems with Applications (ESwA)*, vol. 110, pp. 52–61, 2018.

[97] I. Hernaez, I. Luengo, E. Navas, M. Zubizarreta, I. Gaminde, and J. Sanchez, "The basque speechdat (ii) database: a description and first test recognition results," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Geneva, Switzerland), pp. 1549–1552, International Speech Communication Association (ISCA), 2003.

[98] H. Hoge, H. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach, and K. Choukri, "European speech databases for telephone applications," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Munich, Germany), pp. 1771–1774, Institute of Electrical and Electronics Engineers (IEEE), 1997.

[99] I. Odriozola, I. Hernaez, M. I. Torres, L. J. Rodriguez-Fuentes, M. Penagarikano, and E. Navas, "Basque speecon-like and basque speechdat MDB-600: speech databases for the development of ASR technology for basque," in *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, (Reykjavik, Iceland), pp. 2658–2665, European Language Resource Association (ELRA), 2014.

[100] X. Zalbide, I. Gaminde, I. Hernaez, M. Zubizarreta, and E. Navas, "Euskararako sampa kodeaz," *Euskalingua*, vol. 2, pp. 171–177, 2003.

[101] S. Young, N. Russell, and J. Thornton, "Token Passing: a simple conceptual model for connected speech recognition systems," tech. rep., University of Cambridge, Department of Engineering, 1989.

[102] H. Ney, *Speech recognition and coding, new advances and trends*, ch. Search strategies for Large-Vocabulary Continuous-Speech Recognition, pp. 210–225. NATO ASI Series, 1995.

[103] I. Odriozola, L. Serrano, I. Hernaez, and E. Navas, "The AhoSR automatic speech recognition system," in *Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH)*, (Gran Canaria, Spain), pp. 279–288, 2014.

[104] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[105] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proc. of Asia-Pacific Signal and Information Processing Association - Annual Summit and Conference (APSIPA ASC)*, (Sapporo, Japan), pp. 131–137, 2009.

[106] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, (Waikoloa, USA), pp. 1–4, Institute of Electrical and Electronics Engineers (IEEE), 2011.

[107] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, "The RWTH aachen university open source speech recognition system," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Brighton, United Kingdom), pp. 2111–2114, International Speech Communication Association (ISCA), 2009.

[108] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," tech. rep., Sun Microsystems, 2004.

[109] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to finnish," *Computer Speech & Language*, vol. 20, no. 4, pp. 515–541, 2006.

[110] T. Rotovnik, M. S. Maucec, and Z. Kacic, "Large vocabulary continuous speech recognition of an inflected language using stems and endings," *Speech Communication*, vol. 49, no. 6, pp. 437–452, 2007.

[111] H. Sak, M. Saraclar, and T. Güngör, "Morphology-based and sub-word language modeling for turkish speech recognition.," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Dallas, USA), pp. 5402–5405, Institute of Electrical and Electronics Engineers (IEEE), 2010.

[112] G. F. Choueiter, D. Povery, S. F. Chen, and G. Zweig, "Morpheme-based language modeling for Arabic LVCSR," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Toulouse, France), pp. 1053–1056, Institute of Electrical and Electronics Engineers (IEEE), 2006.

[113] P. Mihajlik, T. Fegyó, Z. Tüske, and P. Ircing, "A morpho-graphemic approach for the recognition of spontaneous speech in agglutinative languages - like Hungarian.," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Antwerp, Belgium), pp. 1497–1500, International Speech Communication Association (ISCA), 2007.

[114] R. Thangarajan, *Modern speech recognition approaches with case studies*, ch. Speech recognition for agglutinative Languages. InTech, 2012.

[115] V. G. Guijarrubia, M. I. Torres, and R. Justo, "Morpheme-based automatic speech recognition of basque," in *Proc. of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, (Póvoa de Varzim, Portugal), pp. 386–393, 2009.

[116] Z. Fang, Z. Guoliang, and S. Zhanjiang, "Comparison of different implementations of mfcc," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[117] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *proc. of IEEE*, (Paris, France), pp. 257–286, Institute of Electrical and Electronics Engineers (IEEE), 1989.

[118] S. J. Young, J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy modelling," in *Proc. of ARPA workshop on Human Language Technology (HLT)*, (Plainsboro, USA), pp. 307–312, 1994.

[119] A. Hunt and S. McGlashan, "Speech recognition grammar specification," tech. rep., World Wide Web Consortium, 2004.

[120] X. Li and Y. Zhao, "A fast and memory-efficient n-gram language model lookup method for large vocabulary continuous speech recognition," *Computer Speech & Language*, vol. 21, no. 1, pp. 1–25, 2007.

[121] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, (Denver, USA), pp. 257–286, 2002.

[122] K. Demuynck, J. Duchateau, D. Van Compernolle, and P. Wambacq, "An efficient search space representation for large vocabulary continuous speech recognition," *Speech Communication*, vol. 30, no. 1, pp. 37–53, 2000.

[123] A. Cardenal, *Realización de un reconocedor de voz en tiempo real para habla continua y grandes vocabularios*. PhD dissertation, University of Vigo, Department of Signal Theory and Communications, 2001.

[124] S. Ortmanns and H. Ney, "Look-ahead techniques for fast beam search," *Computer Speech & Language*, vol. 14, no. 1, pp. 15–32, 2000.

[125] R. Kanters, C. Cucchiarini, and H. Strik, "The Goodness of Pronunciation algorithm: a detailed performance study," in *Proc. of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, (Warwickshire, UK), pp. 2–5, International Speech Communication Association (ISCA), 2009.

[126] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Rhodes, Greece), pp. 2379–2382, International Speech Communication Association (ISCA), 1997.

[127] C. Lopes and F. Perdigao, *Speech technologies*, ch. Phone Recognition on the TIMIT Database, pp. 285–302. Ivo Ipsic, 2011.

[128] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, no. 37, pp. 1641–1648, 1989.

[129] J.-L. Gauvain and L. F. Lamel, "Identification of non-linguistic speech features," in *Proc. of ARPA workshop on Human Language Technology (HLT)*, (Stroudsburg, USA), pp. 96–101, 1993.

[130] T. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, vol. 5, pp. 298–305, 1994.

[131] L. Tóth, "Modeling long temporal contexts in convolutional neural network-based phone recognition," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (South Brisbane, Australia), pp. 4575–4579, Institute of Electrical and Electronics Engineers (IEEE), 2015.

[132] L. Tóth, "Phone recognition with hierarchical convolutional deep maxout networks," *Journal on Audio, Speech and Music Processing*, vol. 2015, p. 25, 2015.

[133] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, USA: Prentice Hall PTR, 1st ed., 2001.

[134] S. M. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[135] I. Luengo, *Análisis y Evaluación de Parámetros para Identificación Automática de Emociones en el Habla*. PhD dissertation, University of the Basque Country, Department of Electronics and Telecommunications, 2010.

[136] I. Odriozola, E. Navas, I. Hernaez, I. Sainz, I. Saratxaga, J. Sánchez, and D. Erro, "Using an ASR database to design a pronunciation evaluation system in basque," in *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, (Istanbul, Turkey), pp. 4122–4126, European Language Resource Association (ELRA), 2012.

[137] I. Odriozola, O. Jokisch, I. Hernáez, and R. Hoffmann, "Diseño y desarrollo de un sistema de evaluación automática de la pronunciación para el euskara," *Procesamiento del Lenguaje Natural*, vol. 49, pp. 101–108, 2012.

[138] B. Mak, M. Siu, M. Ng, Y.-C. Tam, Y.-C. Chan, K.-W. Chan, K.-Y. Leung, S. Ho, F.-H. Chong, J. Wong, and J. Lo, "Plaser: Pronunciation learning via automatic speech recognition," in *Proc. of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing - Volume 2*, (Stroudsburg, USA), pp. 23–29, 2003.

[139] C. for Cultural Co-operation (Education Committee Modern Languages Division), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Applied Linguistics Series, Cambridge University Press, 2001.

[140] P. Adenot, C. Wilson, and C. Rogers, "Web audio api - w3c working draft 10 october 2013," tech. rep., World Wide Web Consortium, 2013.

[141] D. C. Burnett, A. Bergkvist, C. Jennings, and A. Narayanan, "Media capture and streams - w3c last call working draft 14 april 2015," tech. rep., World Wide Web Consortium, 2015.

[142] I. Hickson, "The web sockets api - w3c working draft 22 december 2009," tech. rep., World Wide Web Consortium, 2009.

[143] M. Mustafa, T. Allen, and L. Evett, *Research and Development in Intelligent Systems XXXI*, ch. A Review of Voice Activity Detection Techniques for On-Device Isolated Digit Recognition on Mobile Devices, pp. 317–329. Springer International Publishing, 2014.

[144] T. Virtanen, R. Singh, and B. Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley Publishing, 1st ed., 2012.

[145] I. Mporas, O. Kocsis, T. Ganchev, and N. Fakotakis, "Robust speech interaction in motorcycle environment," *Expert Systems with Applications*, vol. 37, no. 3, pp. 1827–1835, 2010.

[146] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni, and F. Piazza, "An integrated system for voice command recognition and emergency detection based on audio signals," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5668–5683, 2015.

[147] T. Kostoulas, I. Mporas, O. Kocsis, T. Ganchev, N. Katsaounos, J. J. Santamaria, S. Jimenez-Murcia, F. Fernandez-Aranda, and N. Fakotakis, "Affective speech interface in serious games for supporting therapy of mental disorders," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11072–11079, 2012.

[148] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, pp. 250–271, 2017.

[149] T.-W. Kuan, H.-C. Tsai, J.-F. Wang, J.-C. Wang, B.-W. Chen, and Z.-Y. Lin, "A new hybrid and dynamic fusion of multiple experts for intelligent porch system," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9288–9296, 2012.

[150] B. Martínez-González, J. M. Pardo, J. D. Echeverry-Correa, and R. San-Segundo, "Spatial features selection for unsupervised speaker segmentation and clustering," *Expert Systems with Applications*, vol. 73, pp. 27–42, 2017.

[151] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, "New approach in quantification of emotional intensity from the speech signal: emotional temperature," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9554–9564, 2015.

[152] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *Journal on Audio, Speech and Music Processing*, vol. 2015, p. 91, 2015.

[153] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings, Part I: Communications, Speech and Vision*, vol. 4, pp. 377–380, 1992.

[154] V. Hautamäki, M. Tuononen, T. Niemi-Laitinen, and P. Fränti, "Improving speaker verification by periodicity based voice activity detection," in *Proc. of the International Conference on Speech and Computer (SPECOM)*, vol. 2, (Moscow, Russia), pp. 645–650, 2007.

[155] A. Benyassine, "Itu-t recommendation g.729 annex b: a silence compression scheme for use with g.729 optimized for v.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997.

[156] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for german connected digit recognition.," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Budapest, Hungary), International Speech Communication Association (ISCA), 1999.

[157] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Systems Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[158] P. K. Ghosh, A. Tsiartas, and S. S. Narayanan, "Robust voice activity detection using long-term signal variability.," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[159] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *Journal on Audio, Speech and Music Processing*, vol. 2013, no. 1, p. 87, 2013.

[160] K. H. Woo, T. Y. Yang, K. J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronic Letters*, vol. 36, no. 2, 2000.

[161] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics.," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109–118, 2002.

[162] P. Pollak and P. Sovka, "Cepstral speech/pause detectors," in *IEEE Workshop on Nonlinear Signal and Image Processing*, (Halkidiki, Greece), pp. 388–391, 1995.

[163] E. Nemer, R. A. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain.," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[164] S. G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 478–482, 2000.

[165] J. Tatarinov and P. Pollák, "Hmm and ehmm based voice activity detectors and design of testing platform for vad classification," *Digital Technologies*, vol. 1, pp. 1–4, 2008.

[166] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 ibm spine evaluation system," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Orlando, USA), pp. I–53–I–56, Institute of Electrical and Electronics Engineers (IEEE), 2002.

[167] H. Veisi and H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement," *IET Signal Processing*, vol. 6, no. 1, pp. 54–63, 2012.

[168] s. Varela, R. San-Segundo, and L. A. Hernández, "Combining pulse-based features for rejecting far-field speech in a hmm-based voice activity detector," *Computers and Electrical Engineering*, vol. 37, no. 4, pp. 589–600, 2011.

[169] D. Enqing, L. Guizhong, Z. Yatong, and C. Yu, "Voice activity detection based on short-time energy and noise spectrum adaptation," in *Proc. of IEEE International Conference on Signal Processing (ICSP)*, (Beijing, China), p. 464–467, Institute of Electrical and Electronics Engineers (IEEE), 2002.

[170] J. Ramirez, P. Yelamos, J. Gorriz, J. Segura, and L. Garcia, "Speech/non-speech discrimination combining advanced feature extraction and svm learning," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Pittsburgh, USA), pp. 1662—1665, International Speech Communication Association (ISCA), 2006.

[171] J. Ramirez, P. Yelamos, J. M. Gorriz, and J. C. Segura, "SVM-based speech endpoint detection using contextual speech features," *Electronic Letters*, vol. 42, no. 7, pp. 426–428, 2006.

[172] Y. W. Tan, W. J. Liu, W. Jiang, and H. Zheng, "Hybrid svm/hmm architectures for statistical model-based voice activity detection," in *Proc. of International Joint Conference on Neural Networks (IJCNN)*, (Anchorage, USA), pp. 2875–2878, 2014.

[173] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Vancouver, Canada), pp. 7378–7382, Institute of Electrical and Electronics Engineers (IEEE), 2013.

[174] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Florence, Italy), pp. 2519–2523, Institute of Electrical and Electronics Engineers (IEEE), 2014.

[175] Y. Obuchi, "Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression," in *ICASSP*, (Shanghai, China), pp. 5715–5719, Institute of Electrical and Electronics Engineers (IEEE), 2016.

[176] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time Voice Activity Detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.

[177] X. Zhang and J. Wu, "Deep belief networks based Voice Activity Detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 21, pp. 697–710, 2013.

[178] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "Speecon – speech databases for consumer devices: Database specification and validation," in *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, (Las Palmas, Spain), pp. 329–333, European Language Resource Association (ELRA), 2002.

[179] I. Luengo, E. Navas, I. Odriozola, I. Saratxaga, I. Hernáez, I. Sainz, and D. Erro, "Modified LTSE-VAD algorithm for applications requiring reduced silence frame misclassification.," in *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, (Valletta, Malta), pp. 1539–1544, European Language Resource Association (ELRA), 2010.

[180] B. Kotnik, P. Sendorek, S. Astrov, T. Koç, T. Çiloglu, L. Docío Fernández, E. Rodríguez Banga, H. Höge, and Z. Kacic, "Evaluation of voice activity and voicing detection," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Brisbane, Australia), pp. 1642–1645, International Speech Communication Association (ISCA), 2008.

[181] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust dsr front-end on aurora databases," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (Denver, USA), International Speech Communication Association (ISCA), 2002.

[182] J. Ramirez, J. C. Segura, C. Benitez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 3–4, 2004.

[183] M. Westphal, "The use of cepstral means in conversational speech recognition.," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Rhodes, Greece), International Speech Communication Association (ISCA), 1997.

[184] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[185] P. N. Garner, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Communication*, vol. 53, no. 8, pp. 991–1001, 2011.

[186] F.-H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," in *Proc. of ARPA workshop on Human Language Technology (HLT)*, (Stroudsburg, USA), pp. 69–74, 1993.

[187] F.-H. Liu, R. Stern, A. Acero, and P. Moreno, "Efficient cepstral normalization for robust speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Adelaide, Australia), Institute of Electrical and Electronics Engineers (IEEE), 1994.

[188] B. Widrow, R. G. Winter, and R. A. Baxter, "Layered neural nets for pattern recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 110.7, pp. 1109–1118, 1988.

[189] W. H. Delashmit and M. T. Manry, "Recent developments in multilayer perceptron neural networks," in *Proc. of the 7th Annual Memphis Area Engineering and Science Conference (MAESC)*, (Memphis, USA), 2005.

[190] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Volume 1: Foundations* (D. E. Rumelhart, J. L. McClelland, *et al.*, eds.), pp. 318–362, Cambridge: MIT Press, 1987.

[191] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington: Spartan Books, 1962.

[192] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. Cambridge, USA: MIT Press, 1969.

[193] R. Collobert and S. Bengio, "Links between perceptrons, MLPs and SVMs," in *Proc. of International Conference on Machine Learning (ICML)*, (Banff, Canada), 2004.

[194] G. Holmes, A. Donkin, and I. H. Witten, "Weka: a machine learning workbench," in *Proc. of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems*, pp. 357–361, August 1994.

[195] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[196] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

[197] A. Abdulaziz and V. Kepuska, "Noisy timit speech (ldc2017s04)," 3 2017.

[198] I. T. Union, "Generic sound activity detector (gsad); series g: Transmission systems and media, digital systems and networks: Digital terminal equipments-coding of voice and audio signals. g.720.1," tech. rep., Telecommunication standardization sector of ITU (ITU-T), 2010.

[199] I. T. Union, "Coding of speech at 8 kbit/s using conjugate structure algebraic-code-excited linear prediction (cs-acelp); series g: Transmission systems and media, digital systems and networks: Digital terminal equipments-coding of voice and audio signals. g.729," tech. rep., Telecommunication standardization sector of ITU (ITU-T), 2012.

[200] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[201] O. Viikki and K. Laurila, "Noise robust hmm-based speech recognition using segmental cepstral feature vector normalization," in *Robust Speech Recognition for Unknown Communication Channels*, (Pont-à-Mousson, France), pp. 107–110, ISCA, 1997.

[202] S. Tibrewala and H. Hermansky, "Multi-band and adaptation approaches to robust speech recognition," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, (Rhodes, Greece), International Speech Communication Association (ISCA), 1997.

[203] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[204] O. M. Strand and A. Egeberg, "Cepstral mean and variance normalization in the model domain," 2004.

[205] N. V. Prasad and S. Umesh, "Improved cepstral mean and variance normalization using bayesian framework.," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, (Olomouc, Czech Republic), pp. 156–161, Institute of Electrical and Electronics Engineers (IEEE), 2013.

[206] D. Willett, "Online maximum-likelihood mean and variance normalization for speech recognition," August 2015. US Patent App. 14/640,912.

[207] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[208] C.-P. C. Karim, C.-p. Chen, K. Filali, and J. A. Bilmes, "Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, (Denver, USA), pp. 241–244, 2002.

[209] P. Pujol, D. Macho, and C. Nadeu, "On real-time mean-and-variance normalization of speech recognition features," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Toulouse, France), Institute of Electrical and Electronics Engineers (IEEE), 2006.

[210] O. Viikki, D. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise.," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Seattle, USA), pp. 733–736, Institute of Electrical and Electronics Engineers (IEEE), 1998.

[211] M. Ashby and J. Maidment, *Introducing Phonetic Science*. Cambridge Introductions to Language and Linguistics, Cambridge University Press, 2014.

[212] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.