EUSKARAZKO HITZ ISOLATUAK EZAGUTZEKO SISTEMA

HIZKUNTZAREN AZTERKETA ETA PROZESAMENDUA

MASTER OFIZIALEKO
IKASTARO AMAIERAKO LANA

Analysis and Processing of Language
Hizkuntzaren azterketa eta prozesamendua

Igor Odriozola Sustaeta

HAP 2007/2008

I. EDUKIEN AURKIBIDEA:

1. SARRERA

1.1 ASRaren artearen egoera

1.2 Euskararen egoera ASRari dagokionez

1.2.1 Euskarazko ahots-baliabideak

1.2.2 Euskarazko ASR-teknologia

2. HELBURUAK ETA BITARTEKOAK

2.1 Helburuak

2.1.1 Anhitz proiektua

2.2 Bitartekoak

2.2.1 AhoTTS transkribatzailea

2.2.2 HTK softwarea

2.2.3 ZientziaNeteko lexikoia

3. ASR-3200 DATU-BASEA

3.1 ASR-3200 datu-basearen hastapenak

3.2 Datu-basearen osaera

4. EZAGUTZE-SISTEMAREN DISEINUA

4.1 Datuak prestatzea

4.2 Eredu akustikoak entrenatzea

4.3 Testeko ezagutze-prozesua

4.4 Emaitzen analisia

5. PROBEN EMAITZAK

6. ONDORIOAK

7. AIPAMENAK

ERANSKINAK

E.1 ZientziaNeteko lexikoiko lagina

E.2 ASR-3200 datu-baseko datu-fitxategien lagina

E.3 Hasierako HMM prototipoa

II. TAULEN AURKIBIDEA:

2.1 taula AhoLab taldearen transkribatzailearen aukerak

2.2 taula AhoTTS transkribatzailearen «j»ren araua

2.3 taula Euskararako SAMPA ikurrak, adibideekin

3.1 taula ASR-3200eko fitxategietako hizketa motak

4.1 taula Euskararako zehaztutako talde fonetikoak, trifonemak multzokatzeko erabaki-zuhaitzetarako erabiliak

5.1 taula Test-proben xehetasunen laburpena

5.2 taula Test-proben emaitzak (WER), gaussiar kopuruaren arabera

III. IRUDIEN AURKIBIDEA:

1.1 irudia Okerreko hitzen tasak izan duen bilakaera historikoa, hizlariarekiko independente diren sistemetan, gero eta hizketa-estilo konplexuagoetarako

2.1 irudia AnHitz proiektuaren helburuak grafikoki

2.2 irudia AnHitz proiektuko eszenatoki birtualean diseinatuko den bilaketa-sistemaren informazio fluxua

2.3 irudia HTK programa-paketearen software-arkitektura

2.4 irudia HTKren prozesatze-faseak

4.1 irudia Diseinatutako sistema garatzeko urratsen bloke-diagrama

4.2 irudia Fonemak HTK bidez entrenatzeko prozesua

4.3 irudia Gaussiarren nahastura

4.4 irudia Hizketa ezagutzeko prozesuaren eskema orokorra

1. SARRERA

Hizkuntza naturala da gizakiaren komunikazio-tresnarik erabiliena. Hortaz, zentzuzkoa dirudi komunikatzeko mekanismo hori gizakiaren parte-hartzea eskatzen duen prozesu orotara egokitzeak. Azken 40 urteotan, joera horri jarraiki, ahalegin handiak egin dira makinekin hizkuntza naturalen bidez komunikatu ahal izateko; horretarako, TTS sintetizagailuak (*Text To Speech*), eta ASR sistemak (*Automatic Speech Recognition*) garatu dira.

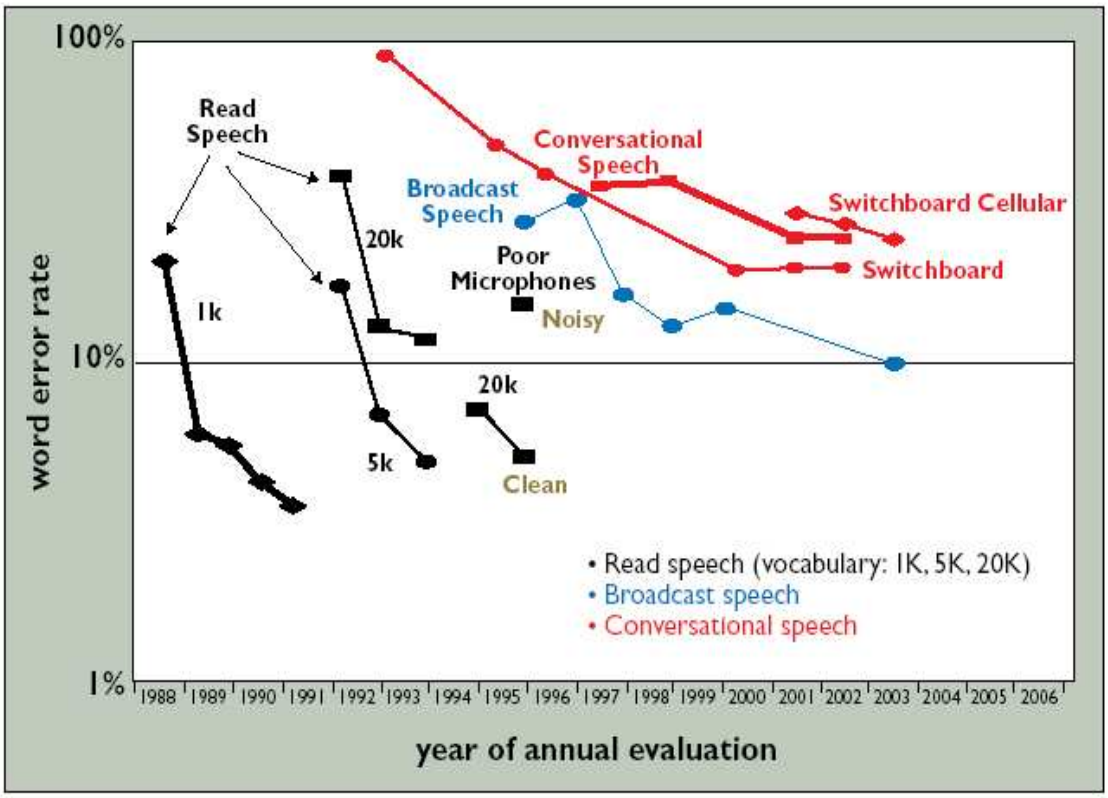
Hala eta guztiz ere, ez da erraza makinei hitz eginaraztea, are zailagoa giza ahotsa ulertaraztea, eta, horretarako, sintesi- eta ezagutze-sistemak osatzeko azpisistema ugari garatu behar izan dira, ahotsa sortzeko eta ulertzeko prozesuei buruzko ikerketa sakonetan oinarrituta. Gaur egun, badira merkatuan TTS eta ASR sistemak oso eraginkorrak direnak, baina, oraindik ere, oso ikerketa-alor garrantzitsuak dira sintetizagailuen ahots-soinuen naturaltasuna eta kalitatea hobetzeko teknikak garatzea eta, halaber, ezagutzaileen ezagutze-prozesuko akats kopurua murrizteko teknikak doitzea [1].

1.1 ASRaren artearen egoera

ASR edo hizketa automatikoki ezagutzeko teknologiak bilakaera nabarmena izan du azken 30 urteetan; denbora-tarte horretan, aurrerapauso handiak eman dira tresna eta arkitektura informatikoei dagokienez, algoritmo estatistiko eragingarriak garatu dira, eta datu kantitate handiak sortu dira. Hala eta guztiz ere, badira zenbait muga oinarrizko eta praktiko, teknologia hori merkatuko alor guztietara hedatzea galarazi dutenak. Hutsune handi bat izan da gizakiaren eta makinaren arteko hizketa-ezagutzan, eta ezin da hizketa ezagutzeko sistemarik erabili, funtzionamendu sendoa eta oso errore gutxikoa izan ezean. Hala, gaur egungo erronkarik handienetako bat izaten jarraitzen du hizketa ezagutzeko sistemaren WER (*Word Error Rate*) edo *okerreko hitzen tasa* murriztea, ingurunea edozein delarik ere [2].

Hizketa ezagutzeko teknologietan diharduten ikertzaileek ahalegin handiak egin dituzte, urte askoan, makinen okerreko hitzen tasak murrizteko. DARPA agentziak hizketa ezagutzeko edo hizketa testu bihurtzeko urtero antolatu ohi dituen ebaluazio-programak dira teknologia horren puntako ikerkuntzaren erakusgarri; 1.1 irudian, 1988 eta 2003 bitartean ebaluazio-programa horietan okerreko hitzen tasan lortutako murrizketa

agertzen da, zenbakitan eta grafikoki adierazirik, hizlariarekiko independenteak diren sistemetarako [3][4]. Ebaluazioetarako, gero eta hizketa konplexuagoak erabiliz joan dira denborak aurrera egin ahala, aurreko mailako hizketarako errore-tasa maila onargarri batera murriztutakoan. Irudiak agertzen du ezen, batez beste, denbora-tarte horretan, urtero % 10eko jaitsiera erlatiboa izan duela errore-murrizketak hizketa-ezagutzan. 1.1 iruditik, halaber, beste bi datu interesgarri ere erauz daitezke: batetik, sistemen emaitzak eta funtzionamendua oso bestelakoak dira, ataza bererako, hizketa-datu garbien eta zaratatsuen (inguruko distortsio akustikoak eragina) artean; desberdintasun horiek ere hauteman dituzte laborategi industrialetan erabiltzen diren ia hizketa-ezagutzaile guztietan, eta ahalegin handiak ari dira egiten alde horiek murrizteko. Bestetik, elkarrizketa-estiloko edo bat-bateko estiloko hizketetan askoz errore gehiago sortzen dira gainerako hizketa-estiloen aldean, ezusteko etenen, ikusizko informaziorik ezaren, ahots-seinalearen heterogeneotasunaren eta beste hamaika zergatiren eraginez. Horiek biek dira gaur egungo hizketa-ezagutzaren teknologiako erronka tekniko nagusiak [2].



1.1 irudia Okerreko hitzen tasak izan duen bilakaera historikoa, hizlariarekiko independente diren sistemetan, gero eta hizketa-estilo konplexuagoetarako.

1. Ingurune akustiko zaratatsuetako hizketa ezagutzeko sistemei dagokienez, oso kaskar jokatzen dute, ataza berean zaratarik gabe ondo funtzionatzen duten sistemek; esaterako, oso ohikoa da telefono mugikor bateko ezagutze-sistema baten errore-tasa onargarria izatea bulego isil batetik dei bat eginez gero, baina onargarria ez izatea telefono mugikor hori bera automobiletik edo aireportu batetik deitzeko erabiltzean. Erronka teknikoa, beraz, honetan oinarritzen da: nola maneiatu aldi bereko zarata gehigarria (hondoko hizlariak barne) eta konboluziozko distortsioa (mikrofono edo hizketa jasotzeko beste gailu merkeago bat erabiltzeagatik sortzen dena). Zaratarekiko sendotasunaren arazoa oso nabaria da akustika aldakorra duten inguruetan.

   Lan handia egin dute ikertzaileek, hizketa-ezagutzaileetan kondizio akustiko aldakorrak karakterizatzeko eta estimatzeko, baita ezagutze-eraginkortasunaren degradazioa eragiten duten iturriak identifikatzeko eta konpentsatzeko ere. Are gehiago, aurrerapauso handiak eman dira, azken urteotan, hizketa ezagutzeko sistemetan, komunitate multimodalean, ikusizko eta entzunezko informazioa integratuz lortutako sendotasun-lorpenak medio.

2. Hizketa-estilo natural eta askeak ezagutzeko sistemei dagokienez, oraingo erronka da gizakiaren pertzepzio-sistemaren ahalik eta antzekoena den sistema bat eraikitzea. Gaur egun, erabiltzaileek, merkatuan dagoen edozein ezagutze-sistemarekin jarduten dutenean, gogoan izan behar dute, uneoro, beren “bikotekidea” makina bat dela; izan ere, makinak oso erraz egiten du huts erabiltzaileek modu askean eta naturalean, lagun batekin egingo luketen bezala, egiten badute berba. Hizketa-ezagutza hainbat alorretara hedatzeko, erabiltzaileak naturaltasunez hitz egiteak ez luke sortu beharko horrenbeste huts.

1.1 irudian ageri denez, DARPAk antolatutako jardunaldietan hizketa-estilo naturaletara eta bat-batekoetara jo dute azken urteotan, eta Japoniako ikerketa asko ere hizketa-estilo naturalerantz bideratu dituzte azkenaldian [5]. Erronka-ildo horren erakusgarri da, orobat, DARPAk 2002an sortutako programa, EARS izenekoa (Effective, Affordable, and Reusable Speech-to-text). Haren helburua zen bost urtean okerreko hitzen tasa urtean % 25 jaistea, 2002ko errore-tasa, % 30-40 ingurukoa, % 5-10era murrizteko.

Oinarri estatistikoari dagokionez, gaur egun, seinale akustikoen eredu akustikoak sortzeko, Markoven Eredu Ezkutuak (HMM, *Hidden Markov Models*) erabiltzen dira gehienbat, eta GMMak (*Gaussian mixture models*) baliatzen dira irteerako probabilitate banaketa adierazteko [6]. Ausazko erregimen geldikorren segida gisa modelatzen da seinalea, eta prozesatzearen lehen fasean, seinale zati laburrak (30 ms ingurukoak) erabiltzen dituzte ASR gehienek, tarte horretan seinalea geldikorra dela jota. Fonemen arteko artikulazioa eta ingurunea modelatzeko, fonemen mugak gainditzen dituzten ereduak erabiltzen dira, eskuarki trifonemak (oinarrizko fonemaren aurreko eta atzeko fonemen testuingurua gordetzen duten ereduak); halere, septafonemak (aurreko hiru eta atzeko hiru fonemen testuingurua gordetzen dutenak) ere erabiltzen ari dira ikertzaileak, puntako ikerkuntza-lanetan [7].

1.2 Euskararen egoera ASRari dagokionez

Euskarazko teknologia ez dago beste ezein hizkuntzaren teknologiatik urrun, baina badu berezitasun bat gainerakoen aldean: euskarak, hiztun kopuru handia ez izateaz eta eremu aski murritz batean mintzatzeaz gainera, alde handiak ditu Euskal Herriko bazter batetik bestera. Hori horrela, bi ikuspuntutatik azter daiteke dibertsitate horrek hizketa-teknologietan duen eragina: fonetikoki eta prosodikoki. Fonetikari dagokionez, zailtasun handiko lana da erabiltzaile guztientzako balioko duen transkribatzaile fonetiko automatiko bat sortzea, eta horrek ondorio esanguratsuak dakartza, bai hizketa ezagutzeko, bai testua ahots bihurtzeko sistemetarako; kontuan izan behar da, orobat, datu-baseak diseinatzean ere sortzen diren zailtasunak. Bigarrenik, prosodiari dagokionez, euskararen eredua ez da bakarra; are gehiago, oso bestelakoa da leku batetik bestera; azentua, esaterako, hizkuntzalariek asko ikertu duten gaia da, baina, batetik, euskararen azentu- eta prosodia-ezaugarri bereziak eta, bestetik, erakunde arau-emaileen erabakitasunik eza direla eta, oraindik ez dago arauturik. Agerikoak dira horrek ezagutzarako eta, batez ere, sintesirako sortzen dituen zailtasunak.

Hizketa-teknologietan, nahitaezkoa da bi alderdi nagusi bereiztea: batetik, ahots-baliabideak; bestetik, teknologia. Ahots-baliabideak sortzea oso prozesu garestia da, eta ezinbesteko tresna dira teknologia behar bezala garatu ahal izateko. Ingeleserako zein alemanerako, datu-base akustiko handiak daude (haietako zenbait askeak, gainera), eta, hortaz, teknologiak ez du oztoporik bere bidean aurrera egiteko. Euskal Herrian, aldiz, oso datu-base akustiko gutxi eta murritzak daude, eta, gainera, ez dira oso erraz

eskuratzekoak. Datu-baseak eraikitzea oso prozesu luze eta garestia da, eta oinarrizko datu-baserik ez izateak, bistan denez, nabarmen baldintzatzen du teknologia behar bezala garatzea.

1.2.1 Euskarazko ahots-baliabideak

Gaur egun, hizkuntza baterako hizketa-teknologiak garatzeko oinarri nagusia ahots-baliabideak dira; hau da, ahotsa eta ahotsari buruzko informazioa, betiere sortu nahi den aplikazioaren araberakoa, daukaten datu-baseak. Oso prozesu garestia eta luzea da ahotsen datu-base handiak ezaugarri jakin batzuen arabera eraikitzea eta audio-seinaleei buruzko informazioa sortzea eta egokitzea. Hortaz, ezinbestekoa da edozein ikerketa-talderentzat baliabideak eskuragarri izatea, ikerketak behar bezala eta modu eraginkorrean gauzatuko badira.

Hizkuntza hedatuetarako, askotariko datu-baseak daude, publikoak zein pribatuak, komertzializatuak zein doakoak, eta horrek erraztu egiten du, hein handi batean, hizketarekin lotutako teknologiak garatzea. Mundu mailako datu-baserik garrantzizkoenak, ahots motakoak nahiz testu motakoak, LDC (*Linguistic Data Consortium*) elkarteak (http://www.ldc.upenn.edu/) eta, Europan, ELRA (*European Language Resources Association*) elkarteak (http://www.elra.info/) komertzializatzen dituzte. Hona hemen euskararako zer datu-base aurki dezakegun, ELRAren webgunera joz gero:

- *Basque FDB-1060* datu-basea [8]: telefono finko bidez grabatutako datu-basea, 1.060 hizlariren grabazioz osatua eta bat-bateko hizketa eta hizketa irakurria dituena. *SpeechDat II* Europako proiektuaren estandarraren arabera eraikia da, eta AhoLab taldeak izan du datu-basea garatzearen ardura.
- *Bizkaifon* datu-basea [9]: bizkaierazko hitz isolatuz, esaldiz, testuz eta herri-literaturako pasartez osatua. Mikrofono bidez grabatutako datu-basea da, eta 21 ordu inguru ditu. AhoLab taldeak Bizkaiko Foru Aldundiaren diru-laguntzaz grabatutako datu-basea da, eta Internet bidez kontsultak egiteko aukera ere badago (http://bizkaifon.ehu.es/).
- *Basque Spoken Corpus, by Jon Aske*: mikrofono bidez grabatutako datu-basea, 42 hizlariren narrazioak dituena.

Aditu programaren web gunean, honako datu-base hauei buruzko informazioa ematen du Eusko Jaurlaritzak:

- *Oinarrizko lexiko fonetikoa*: 60.000 sarrera baino gehiagoko datu-basea, gehien erabiltzen diren hitzak, laburdurak eta akronimoak, eta datu-base akustikoetan bildutako hitzak jasotzen dituena. Transkripzio fonetikoa eta informazio gramatikala ere badu bildua. Eleka enpresak landu du lexikoa.
- *Basque MDB-600*: telefono mugikor bidez grabatutako datu-basea, 660 hizlariren grabazioz osatua. *SpeechDat II* proiektuaren estandarraren arabera dago grabaturik. AhoLab taldeak egina da.
- *ASR3200* datu-base akustikoa: bulego-ingurunean egindako grabazioz osatutako datu-basea, 230 hizlarikoa, hizkuntza-ereduak sortzeko diseinatua. EHUko Zientzia eta Teknologia Fakultateak egina da.

Azken hiru hizkuntza-baliabide horiek, berez, Eusko Jaurlaritzak kudeatzen ditu, baina oraingoz ezin dira eskuratu. Ez dira asko, beraz, euskararako eginda dauden datu-base akustikoak, eta, hortaz, hutsune nabarmena dugu arlo honetan, gainerako hizkuntzen aldean.

Datu-base komertzialez gainera, badira datu-base xumeagoak; alegia, ikerketa-taldeek berek erabiltzeko egiten dituztenak. Eskuarki, ez dira komertzializatzen, baina hitzarmen pribatu batez lor daitezke. Sail honetako datu-baseetan, besteak beste, hauek ditugu:

- Euskarazko corpus fonetikoa: 1995ean EHUren eta UPNAren laguntzaz egina [10] [11].
- *Abiadura* datu-basea (http://aholab.ehu.es/aholab/Resources/Abiadura-Database-4.html): hiru abiaduratan ebakitako esaldiak dituen datu-basea. AhoLab taldeak egina da, eta eskuragarri dago ikerketarako.
- *Hizketa sintetizatzeko ahotsak* (http://aholab.ehu.es/aholab/Resources/Voices-4.html): bi hizlariren ahotsak, emakumezkoarena eta gizonezkoarena, dituen datu-basea, bakoitzeko 702 esaldi fonetikoki orekatu dituena eta 6 emozio eta ebakera neutroa lantzeko aukera ematen duena. AhoLab taldeak egina da, eta eskuragarri dago ikerketarako.

1.2.2 Euskarazko ASR-teknologia

Alor honetan dagoenik eta xede handiena bat-bateko hizketa jarraitua ezagutzea da. Oraindik ez da lortu errore-tasa onargarria duen halako sistemarik, ezta hizkuntza hedatuenetarako ere, nahiz eta emaitzak, pixkanaka, gero eta hobeak izan. Gaur egungo produktu komertzializatuak, nagusiki, hiztegi txiki edo ertaineko aplikazioak dira, hitz segiden aukera murriztuko ataza jakin baterako diseinatuak, teknologia, halako erabileretarako, aski sendoa baita.

Euskarari dagokionez, ez gaude gainerako hizkuntzak baino askoz atzerago; ikerlari askok dihardute ahotsa ezagutzeko aplikazioetan lanean, eta halaxe erakusten dute alor honetako ikerlariek idatzitako hainbat eta hainbat artikuluk eta argitalpenek. Aurrerapauso handiak eman dira, dagoeneko, hizketa jarraituko teknologien garapenean; aipagarria da, esate baterako, Karmele Lopez de Ipiñak egindako tesia, euskarazko hizketa ezagutzeko sistema baten diseinua azaltzen duena [12].

ASRaren helburu gorena bat-bateko hizketa jarraitua ezagutzea da, baina maila xumeagoan ere erabiltzen dira hizketa ezagutzeko sistemak, ataza jakin batera bideraturik. Hiztegia zenbat eta txikiagoa eta hizlariarekiko lotuagoa, orduan eta emaitza hobeak lortzen dira. Hona hemen, sinpleenetik konplexuenera ordenaturik, zenbait adibide esanguratsu:

- Hitz isolatuak ezagutzeko sistemak (edo komando bidezko sistemak): hitz solteak edo hitz-segidak ezagutzeko sistemak dira. Sistemak aukera sorta itxi bat besterik ez du onartzen, eta hortik kanpoko hitz guztiak ez ditu ezagutzen. Elkarrizketa gidatuetarako baliatzen dira, batez ere; adibidez, telefono bidezko elkarrizketa automatikoak egiteko, edo ordenagailu batekin agindu bidez komunikatzeko. Halako aplikazioak egiteko teknologia badago euskararako, eta, baliabideei dagokienez, Basque FDB-1060 eta MDB-600 datu-baseak (telefono bidezko aplikazioetarako) eta ASR-3200 datu-basea (mikrofono bidezko aplikazioetarako) erabil daitezke, betiere euskara estandarrean jarduteko. Dena dela, teknologia eta baliabideak prest izan arren, oso urria da merkatuaren eskaria, eta horrek erabat baldintzatzen du halako sistemen aplikazioa.
- Hizlaria ezagutzeko aplikazioak: alor honetan, ASRko gainerako alorretan ez bezala, hizketarekin bainoago ahotsarekin lotutako teknologia baliatzen da;

alegia, hizlariaren ahotsaren ezaugarriak erabiltzen dira, hizketaren unitate eta azpiunitateak ezagutu beharrean. Hizlarien ahotsen ezaugarriez eredu estatistikoak sortzen dira, eta, prozesamendu estatistikoen bidez, audio-fitxategi grabatu bateko hizlaria edo une jakin batean hizketan ari den pertsona nor den jakiteko aplikazioak sortzen dira.

- Diktatu-aplikazioak: mikrofono bidez esandakoa testu bihurtzeko aplikazioak. Oraindik ez dago garaturik hiztunarekiko askea den eta hiztegi handia (50.000 hitzetik gorakoa) baliatzen duen sistemarik, ezta atzerriko hizkuntzetarako ere; izan ere, errore-tasa onargarria izateko, hiztunen ahotsera egokiturik egon behar dute sistemek. Hiztegi ertain edo txikiko sistemak egiteko, aldiz, prest dago euskarazko teknologia. Datu-baseei dagokienez, *ASR3200* datu-basea aproposa da halako aplikazioetarako.

Sistema horiek guztiak, batik bat, euskara estandarrerako ikertzen dira, baina euskalkiak ezagutzeko sistemak garatzeko lehen urratsak ere eman dira; AhoLab taldeak, esaterako, mendebaldeko euskarazko hitz isolatuak ezagutzeko sistema bat osatu du, Bizkaifon datu-basea oinarri hartuta [13].

Administrazioei dagokienez, 2003an Eusko Jaurlaritzak lankidetza-hitzarmen bat sinatu zuen *Nuance* enpresarekin, euskarazko ahotsa sintetizatzeko (TTS) eta ezagutzeko (ASR) motorrak garatu zitzan. Euskara ezagutzeko *VoCon 3200* sistema garatu zuen Nuance enpresak, baina, gaur egun, proiektua amaitua izanik ere, ez da ageri haren katalogoan.

2. HELBURUAK ETA BITARTEKOAK

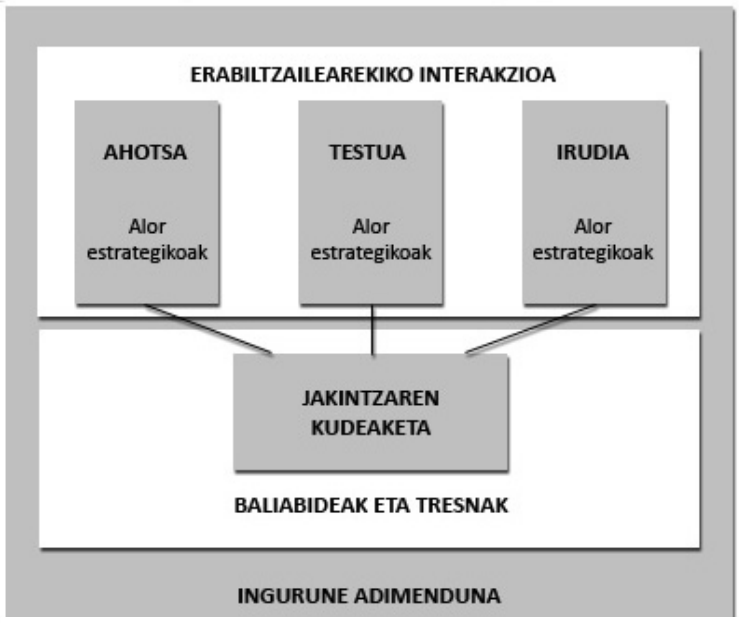
2.1 Helburuak

Lan honen helburu nagusia da euskararako hitz isolatuak ezagutzeko sistema bat aurkeztea. Hitz isolatuak ezagutzeko sistema hori *AnHitz* proiektuaren barnean eraikiko den eszenatoki birtualeko ezagutze-atalean integratuko da, eta beste atal batzuetarako sarrera-datuak eskainiko ditu, datuok beste helburu batzuetarako prozesatzeko.

Hala ere, lan honek baditu beste helburu batzuk ere. Etorkizunera begirako helburu nagusia da hizketa jarraitua ezagutzeko sistema bat eraikitzeko lehen urratsa egitea, eta euskararako hizkuntza-eredu bat osatzeko, eredu akustikoak prest izatea; dena dela, AhoLab laborategiak bere gain hartuko dituen beste zenbait proiektutan ere integratuko da hitz isolatuak ezagutzeko sistema eta, etorkizunera begira, hizketa jarraitua ezagutzeko sistema.

2.1.1 AnHitz proiektua

AnHitz proiektua (http://www.anhitz.com/) Eusko Jaurlaritzaren Industria Sailak Etortek programa baten bidez 2006-2008 aldirako finantzatutako proiektua da, pertsonen eta gailuen arteko interakzioa eta jakintzaren kudeaketa naturala, intuitiboa eta atsegina izan dadin euskarazko hizkuntza-teknologietan ikerketa eta garapena sustatzea helburu duena (ikus 2.1 irudia).



2.1 irudia AnHitz proiektuaren helburuak grafikoki.

AnHitz proiektuak hogeita hamar hilabeteko iraupena du, eta *Hizking 21* (2002-2005) proiektuko partaide berek osatzen dute proiektua. Hauek dira proiektuaren helburu orokorrak: batetik, hizkuntzaren industria garatzeko eta egonkortzeko urrats erabakigarriak ematea, eta, hala, ikerketen emaitza eta gailuak merkatuaren beharretara egokitzea eta produktuen komertzializazioa bideragarri bihurtzea. Bestetik, oinarri teknologikoko enpresa berriak sortzea eta dagoeneko sortuta daudenen egonkortasuna lortzea. Azkenik, jakintza sortzea, ildo estrategiko horretan I+G+Bko erronka eta helburuei autonomiaz erantzuteko ikertzaile talde sendoa prestatuz.

Erakunde hauek hartzen dute parte AnHitz proiektuan: VICOMTech eta Robotiker zentro teknologiek, Elhuyar Fundazioak eta Euskal Herriko Unibertsitateko (UPV-EHU) IXA eta Aholab taldeek.

Aurretik aipatutako lan-eremuen barruan, AnHitz proiektuak helburu zehatzago hauek ditu finkatuta:

- Testu bidez komunikatzea: hizkuntza naturalaren ulermena, hizkuntzaren sorrera eta arrazoitze-sistemak.

- Ezagutzak kudeatzeko sistemak garatzea: galde-erantzun sistemak, informazio-bilaketa eta informazioaren erauzketa aurreratua.

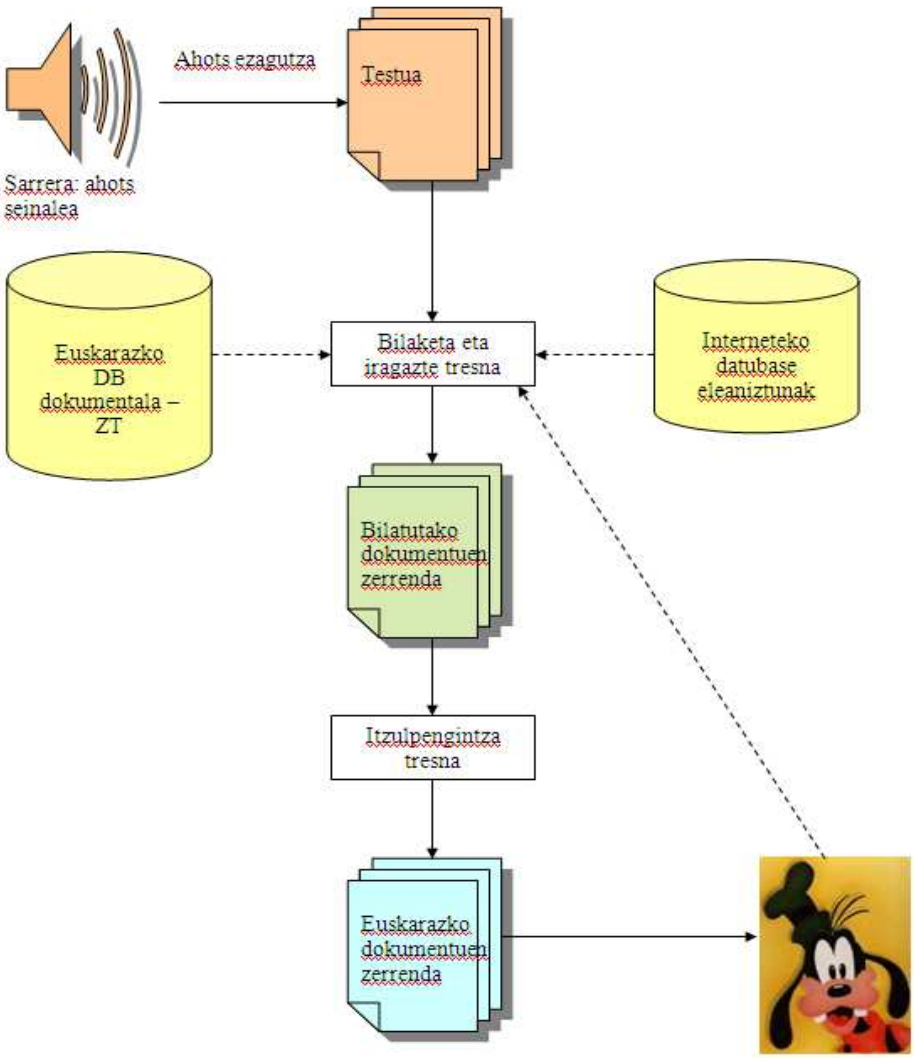
- Elkarrizketa-sistemak: emozioak ezagutzea eta sortzea, elkarrizketa-arauak eta ahotsa sartzea gailu aurreratuetan.

- Itzulpen automatikoa: erabiltzailearekin elkarreraginean aritzeko moduluetan zein ezagutzak kudeatzeko moduluetan aplikatzekoak.

Helburu horiek eszenatoki birtual batean gauzatuko dira, eszenatokian integratuko baita talde guztien ekarpena. Internet bidez erabiltzeko diseinaturik dago eszenatokia, eta edonork erabili ahal izateko prestatuko da. Eszenatokiaren atal nagusiak, oro har, 2.2 irudian ageri dira.

Erabiltzailearen interfazea ahotsa ezagutzeko sistema batez osaturik egongo da; hau da, ahots-seinalea izango da sarrera. Hizketa ezagutzeko teknikak erabiliko dira ahots-seinale hori testu bihurtzeko, eta, ondoren, testua prozesatu egingo da galdegindakoa hainbat datu-basetan bilatzeko, euskarazko datu-baseetan zein Interneteko datu-base eleaniztunetan. Hala, kasurik onenean, erantzun bat aurkitu ostean, euskarara itzuliko da

hala behar baldin bada, eta, prozesuari amaiera emateko, bi modutara emango zaio erantzuna erabiltzaileari: testu gisa eta abatare baten bidez, irudia eta ahots sintetikoa konbinatuta. Hortaz, erabiltzaileak, nahi izanez gero, ez du teklatua erabili beharrik izango galderak egiteko eta erantzunak jasotzeko, galderak ahoz egin ahal izango baititu eta ikusiz eta entzunez jaso.



2.2 irudia AnHitz proiektuko eszenatoki birtualean diseinatuko den bilaketa-sistemaren informazio fluxua

2.2 Bitartekoak

Euskarazko hitz isolatuak ezagutzeko sistema osatzeko, baliabide hauek erabiliko dira:

- ASR-3200 euskarazko datu-base akustikoa: entrenamendurako seinale gisa eta eredu akustiko estatistikoen parametro akustikoak lortzeko

- Transkribatzailea, seinale akustikoen transkripzio fonetikoak izateko.

- HTK softwarea, fonema testuingurudunen eredu akustikoak sortzeko eta entrenatzeko, eta ezagutze-prozesu estatistikoa gauzatzeko.

- ZientziaNet webguneko lexiko-sarrera, ezagutze-sistemaren hiztegia sortzeko.

Atal honetan, transkribatzailearen, HTK softwarearen eta ZientziaNeteko lexiko-sarreraren berri emango da; ASR-3200 datu-basearen xehetasunak 3. atalean azalduko dira, haren garrantzia dela-eta leku berezia behar duelakoan.

2.2.1 AhoTTS Transkibatzailea

Lan honetan erabiliko den transkribatzailea AhoLab laborategiak diseinatutako tresna bat da. Sarrera gisa testu bat sartuz, haren transkripzio fonetikoa, silaben markaketa eta azentuen kokapena itzultzen du. Erabiltzailearen beharretara egokitzearren, hainbat aukera ditu (ikus 2.1 taula).

Aukera	Esanahia
FileRule	Irteerako fitxategia idazteko bidea
HDicDB	Hiztegiaren kokalekuaren bidea
FileOut	Irteerako fitxategiaren izena
Syllable	Silabatan banatzeko aukera ematen du
Stress	Azentuak markatzeko aukera ematen du
Sentence	Testuan agertzen diren puntuazio-ikurrak ateratzeko aukera ematen du.
FileDat	Testuan agertzen diren hainbat daturen zerrenda eskuratzeko aukera ematen du. Hitz, silaba, fonema, azentu eta esaldi kopuruaz gainera, letreiatzen diren hitzen kopuruaren eta ehunekoen gaineko informazioa ere ematen du. Gainera, fonema bakoitzaren informazioa ere ematen du (kopurua eta ehunekoa).

2.1 taula AhoLab taldearen AhoTTS transkribatzailearen aukerak.

AhoLab taldearen transkribatzailea pixkanaka osatuz doan hiztegi batean eta hainbat erregela fonologiko modelatzen dituzten arauetan oinarrituta dago (http://bips.bi.ehu.es/ahoweb/files/otros/reglas_transcripcion_SpeechDateu.pdf). 2.2 taulan, transkribatzaileak erabiltzen dituen zenbait erregela, «j» fonemari dagozkionak, jaso dira, adibide gisa.

Ezaugarria	Hitza	Transkripzio kanonikoa	Transkripzio alternatiboak
Hitz-hasierako «j»	jende	gj e n d e	x e n d e jj e n d e
«n» osteko «j»	laranja	l a r a n gj a	l a r a n x a
Hitz-tarteko «j»	kajoi	k a gj o i	k a x o i

2.2 taula AhoTTS transkribatzailearen «j»ren araua.

AhoTTS transkribatzaileak nazioarteko SAMPA alfabeto fonetikoa darabil. SAMPA (*Speech Assessment Methods Phonetic Alphabet*) makinek irakurri ahal izateko berariaz diseinatutako alfabetoa da (http://www.phon.ucl.ac.uk/home/sampa/). ESPRITen 1541 proiektuaren barnean, SAM (*Speech Assessment Methods*) izenekoan, sortu zuen nazioarteko fonetista talde batek 1987-89 bitartean, eta, hasiera batean, Europako hizkuntza hauetarako soilik diseinatu zen: daniera, nederlandera, ingelesa, frantsesa, alemana eta italiera. Geroago, norvegierarako eta suediarrerako diseinatu zen (1992an), eta, haiei jarraituz, grezierarako, portugeserako eta espainierarako (1993an). BABEL proiektuaren barnean, bulgariera, estoniera, hungariera, poloniera eta errumanierarako osatu zen (1996an). OrienTel proiektuaren barnean, Europatik kanporako jauzia eman, eta arabiarra, hebreera eta turkera gehitu ziren, eta, ildo berari jarraituz, kantonera, kroaziera, txekera, errusiera, esloveniera, japoniera eta koreera ere gehitu ziren.

SAMPAren helburu nagusia zen IPA (*International Phonetic Alphabet*) alfabetoko ikurrak ASCII kodean mapatzea, 33 eta 127 bitartean (7 biteko ASCII karaktere inprimagarriak). IPAko ikurrak ASCIIn mapatzeko gainerako proposamenek ez bezala, SAMPA ez da egile bakar baten eskema, baizik eta hainbat herrialdetako hizketa-teknologien ikertzaileen elkarlanaz sortutakoa. Nazioartean estandarizatuak izan arren, lekuan lekuko hiztunekin garatu eta adostua da, eta, gaur egun, LDCk komertzializatzen dituen datu-base gehienek erabiltzen dute.

Euskarazko SAMPAren eskema 2003an sortu zuen AhoLab taldeak [14], eta, harrezkero, euskararekin jarduten duten ikertzaile askok erabiltzen dute. 2.3 taulan ikus daitezke *SAMPA Basque* eskemaren ikurrak eta fonemak.

Phonetic	Word	Transcription	Description
Plosives			
p	apeza	apes`a	unvoiced bilabial plosive
b	begia	beGia	voiced bilabial plosive
t	etorri	etorri	unvoiced dental plosive
c	ttantta	canca	unvoiced palatal plosive
d	denda	denda	voiced dental plosive
k	ekarri	ekarri	unvoiced velar plosive
g	gaia	gaia	voiced velar plosive
Affricates			
tS	txikia	tSikia	unvoiced palatal affricate
ts	atso	atso	unvoiced apico alveolar affricate
ts`	atzo	ats`o	unvoiced back alveolar affricate
gj	onddo	ongjo	voiced palatal affricate
Fricatives			
jj	leoia	leojja	voiced palatal fricative
f	afaria	afaria	unvoiced labiodental fricative
B	hamabi	amaBi	voiced bilabial approximant
T	perez	pereT	unvoiced interdental fricative
D	adiskide	aDiskiDe	voiced dental approximant
s	hasi	asi	unvoiced apico-alveolar fricative
s`	zoroa	s`oroa	unvoiced back alveolar fricative
S	xoxoa	SoSoa	unvoiced back prepalatal fricative
x	ijito	ixito	unvoiced velar fricative
G	agindua	aGindua	voiced velar fricative approximant
Z	bazkaria	bas`kariZa	voiced prepalatal fricative
Nasals			
m	ama	ama	voiced bilabial nasal
n	neska	neska	voiced alveolar nasal
J	ñabar	JaBarr	voiced palatal nasal
Liquids			
l	lana	lana	voiced alveolar lateral
L	iluna	iLuna	voiced palatal lateral
r	dirua	dirua	voiced alveolar vibrate single
rr	arrunta	arrunta	voiced alveolar vibrate multiple
Front vowels			
i	ipar	iparr	vowel front close unrounded
e	hemen	emen	vowel front mid unrounded
Central vowels			
a	ama	ama	vowel central open unrounded
Back vowels			
o	oso	oso	vowel back mide rounded
u	umore	umore	vowel back close rounded

2.3 taula Euskararako SAMPA ikurrak, adibideekin

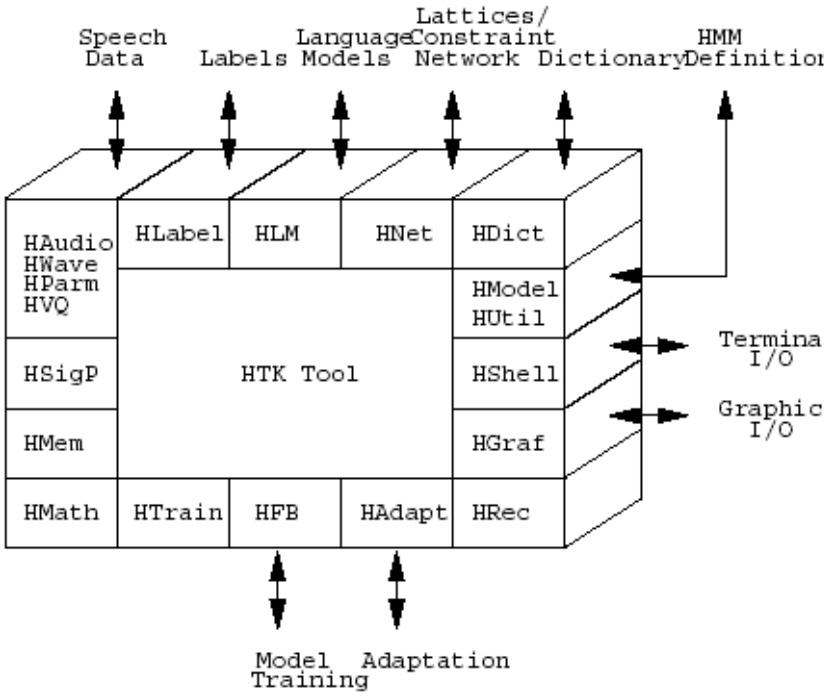
2.2.2 HTK softwarea

HTK softwarea (*Hidden Markov Model Toolkit*) Cambridgeko Unibertsitateak garatutako programa pakete bat da [15] [16]. HMMak eraikitzeko eta maneiatzeko berariaz diseinatutako doako softwarea da, eta, batez ere, ahotsa analizatzeko sistemak (hizketa ezagutzeko, hizlaria ezagutzeko eta segmentatzeko sistemak) diseinatzeko eta entrenatzeko erabiltzen da. Hala ere, beste hainbat aplikaziotarako ere erabiltzen da; besteak beste, hizketa-sintesirako, karaktereak ezagutzeko edo izakion ADNa sekuentziatzeko.

HTK libreria-modulu multzo batez osatua da, eta C lengoaian eskura daiteke; beraz, konpiladore egokia duen edozein makinatan konpila daiteke. Tresnek ezaugarri sofistikatuak dituzte hizketa analizatzeko, HMMak entrenatzeko eta emaitzak analizatzeko. Dentsitate jarraituko gaussiarren nahastura nahiz banaketa diskretua maneiatzen ditu softwareak, eta HMM sistema konplexuak eraikitzeko erail daiteke. HTK banaketak dokumentazio eta adibide ugari ditu, erabiltzaileak urratsez urrats ikasi dezan. HMMei buruzko xehetasun gehiago izateko, jo *HMMen oinarriak* izenburuko eranskinera.

HTK Cambridgeko Unibertsitateko Ingeniaritza Saileko (CUED, *Cambridge University Engeneering Department*) Makina Adimenaren Laborategian sortu eta garatu zuten hastapenetan, Hiztegi handiko hizketa jarraitua ezagutzeko sistemak eraikitzeko (CUED HTK LVR). 1993an, Entropic Research Laboratory Inc. konpainiak HTK saltzeko eskumena erosi zuen, eta, 1995ean, Cambridgeko Entropic ikerketa-laborategiak eskuratu zuen haren eskumen osoa. Handik lau urtera, Entropicek saldu egin zuen HTK, Microsoftek Entropic erosi zuenean. Microsoftek, gaur egun, CUEDri itzuli dio HTKren lizentzia, eta, hala, CUEDek banatu eta, ikertzaile eta boluntarioen laguntzaz, hobetu egiten du HTK, HTK3 web-orriaren bidez. Jatorrizko HTKaren kodearen copyrighta Microsoftek badu ere, ikertzaile eta boluntarioen aldaketa, ekarpen eta zuzenketa berriak HTK3n biltzen dira.

HTKren software-arkitekturari dagokionez, funtzionalitate asko libreria-modulutan eraikia dago; hartara, ziurtatzen da tresna guztiek era berean egiten dutela bitartekaritza-lana kanpoko munduarekin. 2.3 irudian, ohiko HTK banaketaren software-arkitektura ageri da, eta sarrerako eta irteerako interfazeak nabarmendu dira.



2.3 irudia HTK programa-paketearen software-arkitektura

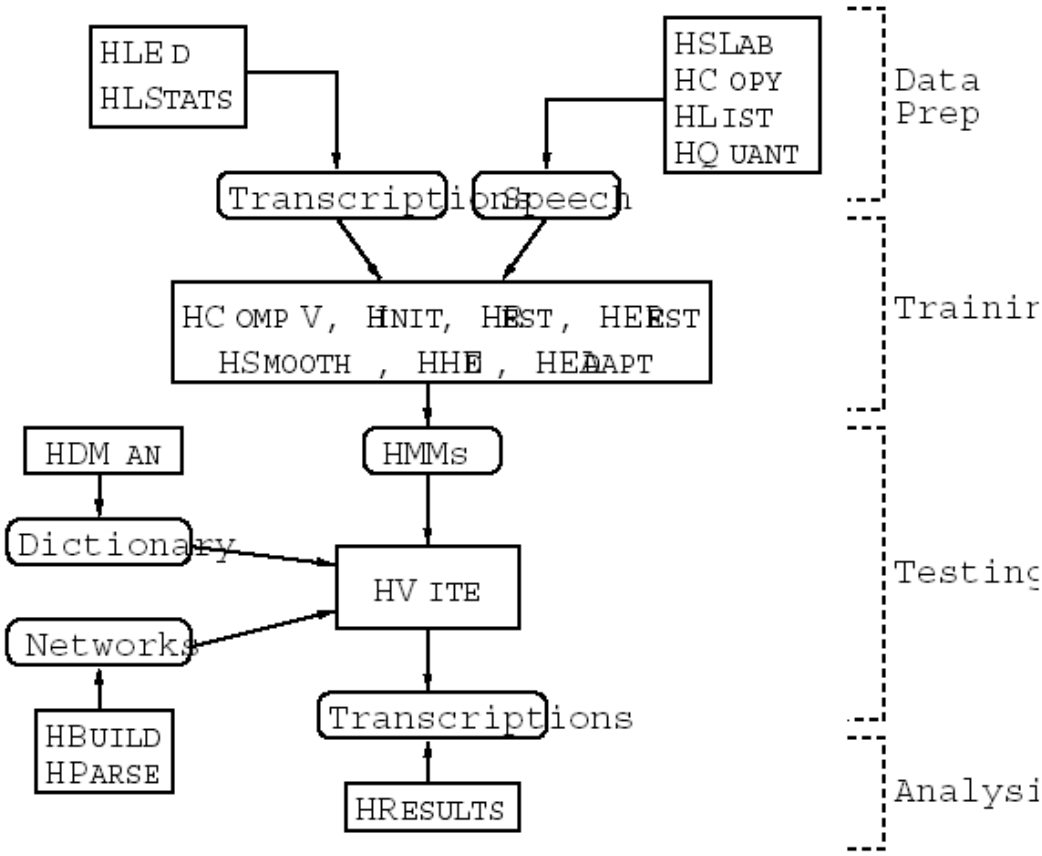
Sarrera eta irteera guztiak eta sistema eragilearekiko truke guztiak HShell libreria-moduluak kontrolatzen ditu, eta memoriaren gorabehera guztiak HMem moduluak hartzen ditu bere gain. Eragiketa matematiko guztiak HMath moduluak maneiatzen ditu, eta audio-seinaleak prozesatzeko beharrezko eragiketak HSigP moduluak kudeatzen ditu. HTK-k erabiltzen duen fitxategi mota bakoitzak berariazko interfaze-modulu bat du. HLabel moduluak etiketa-fitxategiak maneiatzeko interfazea eskaintzen du; HLM moduluak, hizkuntza-moduluen fitxategiak kudeatzeko; HNet moduluak, hitz-sareak eta gramatikak erabiltzeko; HDict moduluak, hiztegiak tratatzeko; HVQ moduluak, VQ *codebook*etarako; eta HModel moduluak, HMMak zehazteko.

Hizketa-seinalearen sarrera eta irteera guztiak, uhin-formaren mailan, HWave moduluaren bidez kudeatzen dira, eta parametrizazio-mailan, HParm moduluaren bidez. Interfaze egonkorra eskaintzeaz gainera, HWave eta HLabel moduluek hainbat fitxategi-formatu maneiatzen dituzte, eta datuak beste sistema batetik inportatzeko aukera ematen dute. Zuzeneko audio-sarrerarako, HAudio modulua erabiltzen da, eta, pantailan grafikoak erakusteko, berriz, HGraf. HUtil moduluak HMMak manipulatzeko hainbat tresna eskaintzen ditu, eta HTrain eta HFB moduluek entrenatzeko tresnak

kudeatzen dituzte. HAdapt modulua adaptazio-tresnez arduratzen da, eta, azkenik, HRec moduluak ezagutze-prozesuaren funtzio nagusiak kudeatzen ditu.

HTK tresnak ohiko komando-lerroetatik erabiltzeko diseinatuta daude. Tresna bakoitzak beharrezko zenbait argumentu dauzka, eta aukerazko beste hainbat. Gainera, libreria-moduluen jokaeraren kontrol zehatza izateko, konfigurazio-aldagaiak ezartzen dira fitxategi batean.

Urrats edo fase finkoak egiteko diseinaturik dago HTK (ikus 2.4 irudia). Lehendabizi, datuak prestatu behar dira, bai hizketa-seinaleak, bai transkripzio ortografiko eta fonetikoak. Horretarako, hainbat tresna ditu HTK-k, datu-base txikiak kudeatzeko, batez ere.



2.4 irudia HTKren prozesatze-faseak

Bigarren fasean, eredu akustikoak sortu eta entrenatu behar dira. Hizketa-seinale segmentaturik izanez gero, hasierako karga-datu gisa erabil daitezke. Kasu horretan, HInit eta HRest tresnek hitz isolatuak entrenatzeko aukera ematen dute. Behar den

HMM bakoitza banan-banan sortzen da. HInit tresnak, entrenatzeko datu guztiak irakurtzen ditu, eta fonema (edo hitz) bakoitzaren lagin guztiak erauzten ditu. Hala, hasierako parametro-balio batzuk sortzen ditu. Lehen zikloan, entrenatzeko datuak uniformeki segmentatzen dira, eredu-egoera bakoitza dagokion datu-segmentuarekin lerrokatzen da, eta, hala, batezbestekoak eta bariantzak kalkulatzen dira. GMM edo gaussiarren nahasturazko ereduak entrenatzen arituz gero, multzokatu egiten dira batezbestekoak. Bigarren zikloan eta hurrengoetan, segmentazio uniformearen ordez Viterbi algoritmoa baliatuz lerrokatzen da. Hala, HInit bidez kalkulatutako hasierako parametro-balioak birkalkulatu egiten dira HRest tresna erabiliz. Horretarako, seinale segmentatuak erabiltzen dira berriro, baina, orain, Baum-Welch-en berrestimazio-prozesua erabiltzen da. Datu segmentaturik izan ezean, *datu lauak* erabil daitezke; hau da, fonemen eredu guztiei balio bera ematen zaie, seinaleen batezbesteko eta bariantza orokorrarena. Horretarako, HCompV tresna erabiltzen da.

Hirugarren fasean, alegia, ereduak sortutakoan, HVite tresna erabiltzen da Viterbiren algoritmoan oinarrituta hizketa ezagutzeko, *token passing* algoritmoa baliatuz. HVitek fitxategi hauek hartzen ditu sarrera gisa: hitz segidak deskribatzen dituen hitz-sare edo gramatika bat, hitz bakoitzaren ebakera zehazten duen hiztegi bat eta HMM multzo bat. Hitz-sarea fonema-mailara itzultzen du, eta, gero, dagokion HMM elementua esleitzen dio fonema bakoitzari. Horrenbestez, hizketa ezagutzeko prozesua bi modutara gauzatu daiteke: memorian gordetako audio-fitxategiak erabiliz eta zuzeneko audio-sarrera erabiliz. Gainera, HVitek hainbat *token* erabiltzeko aukera ematen du, hainbat hipotesi dituen sare bat sortzeko.

Laugarren fasean, ezagutze-sistemaren analisia egiten da. Horretarako, aurrez grabatutako test-seinaleak erabiltzen dira, eta ezagutze-prozesuak emandako irteerak erreferentziazko transkripzioekin alderatzen dira. Konparatze hori HResults izeneko tresnarekin egiten da. Tresna horrek dinamikoki lerrokatzen ditu transkripzio-pareak, eta, hala, ordezkatze-, ezabatze- eta txertatze-erroreak zenbatzen ditu. HResults tresnaren irteerako formatua bateragarria da NIST (US National Institute of Standards and Technology) erakundeak erabiltzen duenarekin, eta, funtzionamenduaren neurri orokorrak ematez gainera, hizlarikako akatsak, nahasmen-matrizeak eta denboraren arabera lerrokatutako transkripzioak ere sortzen ditu.

HTK programak erabiltzearen abantailen artean, hiru nabarmendu daitezke:

- HMMei lotutako algoritmoak diseinatzea saihesten da.
- Banaketa askekoak dira, eta, beraz, doakoak.
- Iturburu-kode gisa banatzen denez, programak aldatzeko eta egokitzeko aukera dago, hala beharko balitz.

Ereduak entrenatzeko eta ezagutze-sistemak diseinatzeko programa sorta bat dira HTK programak. Oso tresna malgua da, eta hona hemen zer aukera dituen:

- Hitzen ereduak edo hitzak baino unitate txikiagoen ereduak entrenatzea.
- Fonema-ereduak entrenatzen diren kasuan, testuingurua erabil dezake (trifonemak, pentafonemak, heptafonemak...).
- Trifonemak erabiltzen direnean, ereduak edo egoerak taldekatzeko aukera ematen du, bai *bottom-up* eran, bai *top-down* eran.
- Ereduen Gaussen osagai kopurua hauta daiteke.
- Transkripzio lerrokatu nahiz ez-lerrokatuen bidez entrenatzea.
- Eredu jarraitu edo diskretuak entrenatzea.
- Hainbat motatan parametrizatzeko aukera.
- Hiztegiak aldatzeko, moldatzeko eta erabiltzeko aukera.

Malgutasun horrek, baina, badu bere ordaina. Programa kopurua oso handia da, eta zuzeneko ordenan exekutatu behar dira, emaitza behar bezalakoa izango bada. Gainera, programa bakoitzean, parametro asko maneiatu behar ditu erabiltzaileak, eta, horretarako, sakon ezagutu beharra dago HTK programen funtzionamendua. Horren guztiaren ondorioz, nahiko konplexua da HTK erabiltzea gutxieneko eskarmentua ez duten pertsonentzat.

3.4 bertsioa da oraingoz lor daitekeen azken bertsioa.

2.2.3 ZientziaNeteko lexikoia

AnHitz proiektuaren barnean diseinatuko den eszenatoki birtualean, pentsatu zen Elhuyar Fundazioaren ZientziaNet webgunearen (http://www.zientzia.net) historian jasotako sarrerak baliatzea, erabiltzailearen ahotsa ezagutzeko sistemarako lexikoi gisa. Sarrerako hitz horiek webguneko bilaketa-sisteman erabiltzaileak egindako galderetatik erauziak dira, eta lexikoiak hitz bakoitzaren maiztasunaren berri ere ematen du. Hitz-zerrendak, beraz, kutsu teknikoa izango du, baina egokia iritzi diote AnHitz proiektuko partaideek.

Kontuan izan beharra dago zerrenda horretan hitz desberdin guztiak bildu direla; hau da, erabiltzaileak gaizki idatzitako, akats ortografikoz idatzitako eta jolasean sartutako hitz ugari daude. Maiztasun handieneko hitzetik maiztasun txikienera ordenatuta daude, eta zentzuzkoa da pentsatzea maiztasun handieneko hitzak direla baliagarrienak; behin baino agertzen ez diren hitz asko akastunak dira.

Lexikoiaren lagin bat jaso da E.1 eranskinean. Han, zerrendaren hasiera eta amaiera ageri dira, lexiko-desberdintasunari hobeto erreparatzeko. Lexikoiak 52.761 sarrera ditu; lehen 69 hitzek (lexikoiaren % 0,13k) mila agerpen baino gehiagoko maiztasuna dute, baina agerpen bakar bateko hitzak 25.310 dira (lexikoiaren % 47,97).

3. ASR-3200 DATU-BASEA

3.1 ASR-3200 datu-basearen hastapenak

ASR-3200 datu-base akustikoa 2003an egina da. Eusko Jaurlaritzaren Hizkuntza Politikarako Sailburuordetzak, 2003an, ahots-teknologiak gizartearen oinarrizko zutabe gisa ikusita, euskarazko ahotsa ezagutzeko eta sintetizatzeko motorrak garatzeko proposamena egin zuen «Euskararen garapen teknolinguistikoa» izeneko txostenaren bidez, eta, Kultura Saileko ordezkariek, Informatika eta Telekomunikazio Zuzendaritzak, Administrazioa Modernizatzeko Bulegoaren Zuzendaritzak eta EJIE Informatika Elkarteek osatutako batzordeak proposamen hori onartu ondoren, Eusko Jaurlaritzak lankidetza-hitzarmena sinatu zuen *Nuance* —lehen *Scansoft Belgium* eta, lehenago, *Lernout & Hauspie*— enpresarekin, euskarazko ahotsa sintetizatzeko (TTS) eta ezagutzeko (ASR) motorrak garatu zitzan. Hizkuntza Politikarako Sailburuordetzaren txostenak zioenez, motor horiek erabilita hainbat aplikazio sortzea zen helburua: arreta-telefonoak, telefono bidezko merkataritza- eta banketxe-eragiketak, herritarrentzako arreta-zerbitzuak, jostailuak, euskarazko irakaskuntzarako aplikazioak, telefono-aurkitegia edota garraio publikorako aplikazioak. Horretarako, euskara ezagutzeko *VoCon 3200* sistema garatu zuen Nuance enpresak, baina, gaur egun, proiektua amaitua izanik ere, ez da ageri haren katalogoan.

Eusko Jaurlaritzak euskal agente teknologikoei (enpresei, unibertsitateei eta hizkuntza-baliabideak garatzen dituztenei) agindu zien euskarazko ahotsa ezagutu eta sintesia egiteko motoreak garatzeko beharrezko hizkuntza-baliabideak lantzea:

- "25 milioi hitzeko testu-corpusa". Hizkuntzaren erabileraren ahalik eta laginik errealistena osatzen duen testu-multzo egituratua da corpusa. Hizkuntzalari eta informatikari talde batek osatua, ADUR SOFTWARE PRODUCTIONS enpresak landu zuen hizkuntza-baliabide hori. Corpusa lortzeko hainbat iturritara jo zen: aldizkari ofizialetara, argitaletxeetara, aldizkarietara, Internetera eta abarretara.

- Oinarrizko lexiko fonetikoa: 60.000 sarrera baino gehiago ditu, eta gehien erabiltzen diren hitzak, laburdurak eta akronimoak eta datu-base akustikoetan bildutako hitzak jasotzen ditu. Hitzen transkripzio fonetikoa gaineratzen da, baita informazio gramatikala ere. ELEKA enpresak landu zuen lexikoa.

- Telefoniarako datu-base akustikoa: hizkuntza-ereduak sortzeko datuak eskuratzeko telefono bidez egindako grabazioak. Datu-basea Euskal Herriko Unibertsitateko Aholab taldeak landu zuen (SpeechDat_eu).

- ASR-3200erako datu-base akustikoa: hizkuntza-ereduak sortzeko datuak eskuratzeko bulego-ingurunean egindako grabazioak. Euskal Herriko Unibertsitateko Zientzia eta Teknologia Fakultateak egin zuen lan hori.

Hizkuntza-baliabide horiek Eusko Jaurlaritzak kudeatzen ditu gaur egun.

Eusko Jaurlaritzaren webgunean, Aditu programaren berri ematen da labur-labur. Programa horren helburua da euskarazko ahotsa ezagutu eta sintesia egiteko motoreak garatzea eta ezartzea. Motore horietan oinarrituta, askotariko aplikazioak garatzeko aukera eskaintzen die Eusko Jaurlaritzak Euskal Autonomia Erkidegoko enpresa eta administrazioei, beren zerbitzuak eta produktuak ahots-teknologiaren bitartez euskaraz ere eskaini ahal izateko. Hala, teknologia horretaz balia daitezkeen eta interesa duten enpresei eta erakundeei bideragarritasun-txostenak egiteko erraztasunak eskaintzen dute. Horrekin, lortu nahi da euskara hizkuntzaren teknologien esparruan finkatzea eta munduko hizkuntzarik mintzatuenen mailan kokatzea. Gainera, euskara Informazioaren Gizartean benetan hedatzeko bitarteako sor daitezen sustatu nahi da.

Ildo beretik, ahotsa ezagutzeko VoCon 3200 sistema enkaptsulatuaren berri ematen du Eusko Jaurlaritzak Adituren webgunean:

«***VoCon 3200.*** *Ahotsa ezagutzeko azken belaunaldiko motorea da, doitasun handikoa eta hiztegi aberatsekoa, zereginik konplexuenak prozesatu ahal izateko. VoCon 3200 sistemak gaur egun Speech2Go™ eta ASR-3200 gisa ezagutzen diren bi produktuen teknologiak biltzen ditu. VoCon 3200 sistema Nuance-ren ASR garapen-sistema barne hartuarekin batera dator. Eskalan alda daitekeen inplementazio-soluzio bat da, programazio azkarrekoa, eta oso baliagarria da softwareko eta hardwareko aplikazioetan ahotsa ezagutzeko aukera txertatzeko. Garapen-tresnen multzoari esker, garatzaileek —ASR sistemako adituek zein teknologia hori erabiltzen hasi diren profesionalek— aplikazio eraginkorrak eta zehatzak sor ditzakete ahotsa ezagutzeko*.»

SpeechDat_eu telefoniarako datu-base akustikoa ELRAren webgunean salgai badago ere, ASR-3200 bulegorako datu-base akustikoa ez dago oraindik salgai, Eusko Jaurlaritzak ez baititu amaitu hura saltzeko egin beharreko gestioak.

3.2 Datu-basearen osaera

ASR-3200 datu-base akustikoa bulego-ingurunean egindako grabazioz osatutako datu-basea da. 230 hizlariren grabazioak ditu, eta hizkuntza-ereduak sortzeko diseinaturik dago. ASR-3200 *Speecon* (Speech-driven Interfaces for Consumer Devices) partzuergoaren datu-base akustikoak sortzeko zehaztapenetan oinarrituta dago (http://www.speechdat.org/speecon/index.html), baina baditu aldeak harekin konparatuz gero.

Speecon proiektuaren helburu nagusia da ahots-bidez gobernatutako interfazeak sortzea erabiltzaile-mailako aplikazioetarako, hainbat hizkuntzatarako eta proiektuan tratatutako ingurune akustikoetarako. Merkatuaren eskariaren arabera hautatu ziren hizkuntzak (18 guztira), eta, ezagutzaileak hizkuntza horietara guztietara eta ingurune akustiko jakin batera moldatzeko, SLR (*Spoken Language Resources*) datu-baseak sortu ziren; hala, SLRekin entrenatzen dira ezagutzaileen eredu akustikoak.

Speecon Proiektuaren barnean, hainbat datu-base akustiko sortzeaz gainera, ohiko ingurune akustikoak hartu ziren kontuan: etxe edo bulegoa, automobila eta leku publikoak. Datu-baseen zehaztapenak ahots bidez gobernatutako interfazeen funtzionaltasunak kontuan izanda diseinatu ziren. Funtzionalitate horien adibide dira komando-ezagutzea, zenbakiak, izenak, datak eta abar. Gainera, erabilera-eremua zabaltzearren, datu-baseak beste ingurune akustiko batzuetara egokitzeko teknikak ere garatu ziren, eta, hala, hizketa ezagutzeko sistemaren funtzionamenduaren erroreak murrizteko aukera dago, berariazko SLR batez entrenatu ez den inguruetarako.

Euskarazko ASR-3200 datu-basearen osaera, esan bezala, Speeconen zehaztapenetan oinarrituta dago, baina baditu zenbait desberdintasun. Hizlari kopurua eta kontuan izandako ingurune akustikoak dira nabarmenenak; izan ere, ASR-3200 datu-basean 230 hizlariren grabazioak biltzen dira, bulego- edo etxe-ingurunean grabatuak; Speeconen zehaztapenak, ostera, azaltzen du 550 hizlari heldu eta 50 ume izan behar direla. Ingurune akustikoei dagokienez, ASR-3200 bulegoan soilik grabatua da; Speeconen zehaztapenak dio, berriz, bulegoan, automobilean eta leku publikoetan grabatuak izan behar dutela.

Hizlari bakoitzaren grabazio guztiak sesio batean gordetzen dira. ASR-3200ek, beraz, 230 sesio ditu, eta, sesio bakoitzean, 316 audio-fitxategi daude (Speeconen arabera, 600 izan behar dute). Fitxategiaren izenean, sesio-zenbakia eta fitxategi-zenbakia agertzen dira, formatu honetan:

U0<SES><FITX>.wav

U: *Utterance* esan nahi du.

0: Fitxategi guztietan agertzen da.

<SES>: Hiru zifrako kodea, sesio-zenbakia adierazten duena (001-230 bitartekoa).

<FITX>: Hiru zifrako kodea, fitxategi-zenbakia adierazten duena (001-316 bit.).

Bestalde, ASR-3200 datu-basean, sesioek datu-fitxategi bana dute, eta fitxategi bakar batean biltzen dira inkestatzaileak egindako galderak, hizlariaren erantzunen transkripzio ortografikoak, grabazioaren nondik norakoak, mikrofonoen ezaugarriak, lekuari eta hizlariari buruzko xehetasunak eta abar. Speeconen zehaztapenen arabera, ordea, fitxategi bakoitzak bere datu-fitxategia izan behar du. Hala ere, sesio bakoitzean inkestatzaileek egindako galderak oso antzekoak dira bi datu-baseetan.

ASR-3200eko fitxategi bakoitzean agertzen diren datuak, oro har, hauek dira:

- *Promptsheet* edo inkesta-orria egiteko erabilitako script-zenbakia.
- Audioa kodetzeko teknika: PCM lineala.
- Maiztasuna: 16.000
- Kanal kopurua: 2
- Byte-formatua: 01
- Karaktere-formatua: ANSI
- Mikrofonoei buruzko xehetasunak eta kokalekua: CLOSE_TALK eta DESKTOP
- Hizlariari buruzko datuak: izen-abizenak, adina, sexua, jaiotze-lekua, euskara-maila.
- Grabazio-lekuari buruzko datuak: kokalekua, gelaren tamaina eta hizlariaren kokapena.
- Grabazioei buruzko datuak: kopuru osoa, gordetzeko direktorioaren izena, grabazio-data, ordua eta iraupena.
- 316 galdera (edo irakurri beharreko testuak).
- 316 galdera horien erantzunen transkripzio ortogratikoak.

Xehetasun gehiago nahi izanez gero, E.2 eranskinean 001 sesioko datu-fitxategiaren zati bat jaso da lagin gisa.

Sesio bakoitzeko 316 fitxategietan, 3.1 taulan agertzen diren datuak daude grabatuta. Taulan, ezkerretik eskuinera, ASR-3200eko fitxategi-zenbakiak, kodeak, esanahia eta hizketa mota horri dagokion Speeconeko kode estandarra ageri dira.

fitxategi-zenbakia	**kodea** (transkripzio bakoitzaren ondoan)	**esanahia**	**Speecon-eko kode estandarra**
CALIBRATION DATA			
001	SY1	Giroko zarata	_01
ez dago		“silence word”	N01
FREE SPONTANEOUS ITEMS			
002-006	SS1-SS6	Free spontaneous items	F01-F30
ELICITED SPONTANEOUS ITEMS			
007-012	T1-T5	Orduak eta datak	ED1-ED2 ET1-ET2
013-017	IWn1-IWn5	Herri-, pertsona-, enpresa-izen bereziak	EP1-EP3 EC1-EC2
018	CA1	Letrak	EL1
019-020	IWa1-IWa2	Bai/Ez	EQ1-EQ2
021	IWn6	Hizkuntza-izen berezia	EO1
022-024	FN1-FN3	Telefono-zenbakia	EN1-EN3
READ SPEECH			
025-064	ISp1-ISp40	40 phonetically rich sentences	S01-S30
065-072	IWp1-IWp8	8 phonetically rich words	W01-W05
CORE WORDS			
073-077	ID1-ID5	Zenbaki solteak	CI1-CI4
078-082	CD1-CD5	Zenbaki-kateak	CB1 CC1-CC4
083	FN4	Telefono-zenbakia (zenbakia irakurria)	CE1
084-086	IN1-IN4	Zenbaki luzeak (irakurria)	CN1-3
087	FN5	Diru kantitate bat (irakurria)	CM1
088-093	T6-T11	Orduak eta datak (irakurria)	CT1-CT2 CD1-CD3
094-097	CA2-CA5	Letrak (irakurria)	CL1-CL3
098-100	IWn7-IWn9	Izen bereziak (irakurria)	CP1 CO1-CO2
101-102	IWa3-IWa4	Bai/Ez (irakurria)	CQ1-CQ2
103-104	IWn10-IWn11	Web helbideak	CW1, CW2
105-316	IWa5-IWa216	Komandoak	Y01-Y99 101-185 … 901-995

3.1 taula ASR-3200eko fitxategietako hizketa motak

Hortaz, sesio bakoitzeko 001-024 fitxategietan, bat-bateko hizketa jasotzen da, eta, gainerakoetan, irakurria.

Datu-fitxategietan, zenbait etiketa berezi agertzen dira transkripzio ortografikoetan txertatuta. 3.2 taulan, etiketa horiek eta bakoitzaren azalpena bildu dira.

Etiketa	Azalpena
FRA	Hitz zatitu bat adierazten du
FIL	Hizlariaren etenune-betegarri bat adierazten du
INT	Aldizka errepikatzen den zarata adierazten du
SPK	Hizlariaren zarata bat adierazten du.
UNI	Ulertzen ez den hitz bat adierazten du.
TRC	Fitxategiaren hasieran edo amaieran hitz bat moztuta dagoela adierazten du.

3.2 taula ASR-3200eko transkripzioetako etiketa bereziak eta haien azalpena

Bestalde, hizlarien ezaugarriei dagokienez, 230 hizlarien artean 127 (% 55,22) emakumezkoak dira; 103, berriz, gizonezkoak (% 44,78).

Hizlariek darabilten euskarari dagokionez, betiere 001-024 fitxategiak kontuan hartuta —gainerakoak irakurriak baitira—, honela banatzen dira:

- Euskara batua: 111 hizlarik (% 48,26k).
- Euskara batu kutsatua: 41 hizlarik (% 17,83k). 14k (% 6,09k) ekialdeko ukituak dituzte, eta 27k (% 11,74k) mendebaldekoak.
- Nahastea: 10 hizlarik (% 4,35ek). Mendebaldeko hizkerak eta batua nahasirik.
- Euskalki hutsa: 68k (% 29,56k). Oro har, 47k (% 20,43k) mendebaldekoa egiten dute, eta 21ek (% 9,13k) ekialdekoa.

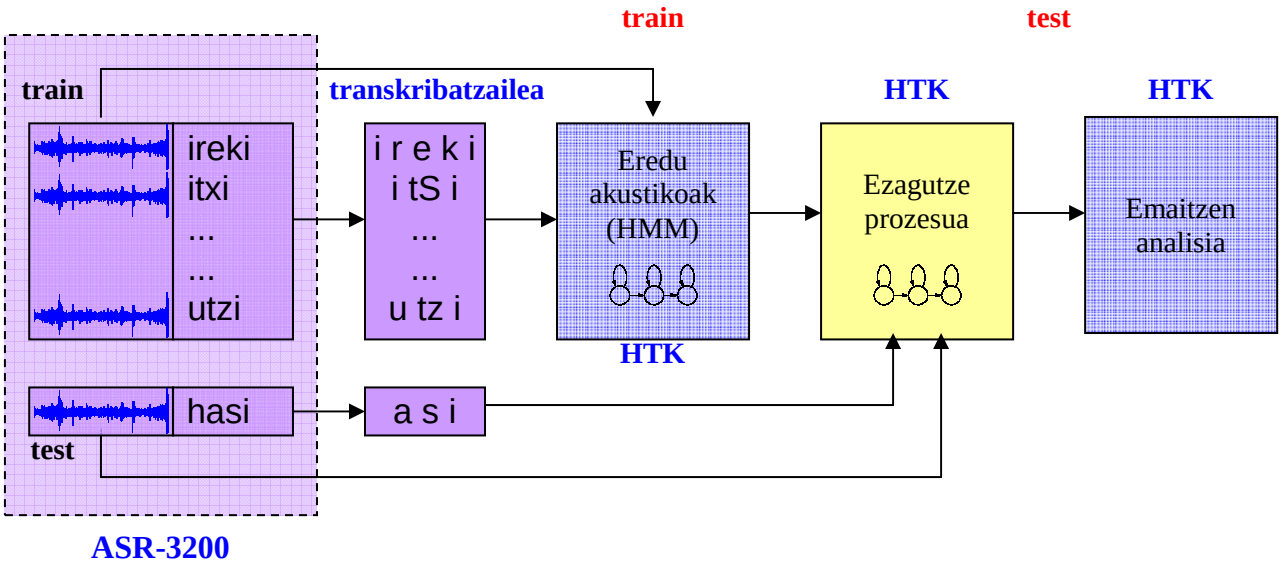
Aipagarria da, gainera, mendebaldeko zenbait hizkeretako Z soinu bereizgarria (frikari aurresabaikari ahostuna) 10 hizlarik erabiltzen dutela.

Hori guztia kontuan izatea ezinbestekoa izango da, beraz, transkripzio fonetikoak egitean.

4. EZAGUTZE-SISTEMAREN DISEINUA

Euskarazko hitz isolatuak ezagutzeko sistema diseinatzeko, ezagutze-sistema generiko bat diseinatzeko lau urrats orokorrei jarraitu zaie: datuak prestatzea, eredu akustikoak entrenatzea, ezagutze-prozesua egitea eta emaitzen analisia.

4.1 irudian, sistema garatzeko bete diren urratsen bloke-diagrama ageri da.



4.1 irudia Diseinatutako sistema garatzeko urratsen bloke-diagrama

4.1 Datuak prestatzea

Hitz isolatuak ezagutzeko sistemaren funtzionamendu ebaluatu ahal izan dadin, ASR-3200 datu basea bi azpicorpusetan banatu da: *train* atala, eredu akustikoak entrenatzeko balioko duena; eta *test* atala, eredu akustiko horiez gauzatutako ezagutze-prozesua ebaluatzeko balioko duena. Train azpicorpuserako, datu-basearen bi heren hautatu dira; hortaz, 155 sesiok (edo hizlarik) osatzen dute train azpicorpusa eta 75 sesiok test azpicorpusa.

Gainera, AnHitz proiektuko eszenatoki birtualerako hala behar delakoan, hizketa irakurria duten fitxategiak baino ez dira hautatu; hau da, sesio bakoitzeko 025-316 fitxategiak. Hortaz, 45.260 (155 ses × 292 fitx/ses.) fitxategi ziren, hasiera batean, baliagarriak ereduak entrenatzeko. Hala eta guztiz ere, kendu egin dira transkripzioetan etiketa bereziak dituzten fitxategiak (ikus 3.2 taula), ahalik eta eredu garbienak sortzearren, eta, hala, 45.000 fitxategi erabili dira guztira ereduak sortu eta entrenatzeko.

ASR-3200 datu-baseak, 3. atalean aipatu den bezala, seinaleak ditu batetik, eta hizketa-seinale horien transkripzio ortografikoak bestetik. Transkripzio ortografikoak moldatu egin behar izan dira eskuz, zenbait oztopo aurkitu baitira transkripzio ortografiko horien baliokide fonetikoak sortu ahal izateko. Hona hemen zenbait adibide:

- Zenbait transkripzio ortografikoki automatikoki sortuak dira, hizlariak modu bakar batean esango dituelakoan. Hori ez da, baina, beti hala gertatzen; adibidez, «www» irakurtzeko agindutakoan, «doble uve doble uve doble uve» irakurtzen dute zenbaitek, «hiru uve doble» beste zenbaitek, «uve bikoitz hirukoitza» beste zenbaitek, «hiru uve bikoitza» beste zenbaitek eta «hiru uve bikoitz» beste zenbaitek. Horietan guztietan agertzen den transkripzioa, ordea, «hiru uve bikoitz» da.

- «@» irakurtzeko agindutakoan, zenbaitek «a bildua» esaten dute, beste batzuek, ostera, «arroba». Transkripzio ortografikoetan, berriz, «a bildua» ageri da.

Transkripzio fonetikoak osatzeko, AhoTTS transkribatzailea erabili da, eta euskararako SAMPA alfabetuko zenbait ikur batu egin dira, testuinguruak eragindako alofonoak direnean; izan ere, fonema testuingurudunak (trifonemak) erabili dira, eta, beraz, trifonemek gai izan beharko lukete, esaterako, /g/ fonema hurbilkari denean eta ez denean hautemateko. Hala, bada, ebakera hurbilkaria duten fonemak ikur bakar batean batu dira. Mendebaldeko hizkeretako /Z/ frikari aurresabaikari ahostuna ere ez da kontuan hartu, hizketa irakurrian ez baita agertzen. Hortaz, 30 ikur erabili dira guztira.

Bestalde, zenbait zuzenketa ere egin behar izan dira, eskuz, hizlariek ez baitute dena espero bezala esaten eta ebakitzen. Esaterako:

- Komandoen atalean (ikus 3.1 taula), ordenagailuarekin zerikusia duten siglak agertzen dira, eta siglok hainbat modutara esaten dituzte hizlariek. Esaterako: «MP3» irakurtzeko eskatzen zaienean (hala da transkripzio ortografikoa ere), «eme pe hiru» esaten dute batzuek, eta «eme pe tres» beste batzuek. «DAT» ere «dat» irakurtzen dute zenbaitek, eta «de a te» beste batzuek. Sigla guztien ebakerak zuzendu dira.

- Transkribatzaileak «j» ortografikoak /gj/ bihurtzen ditu hitz hasieran edo sudurkari edo dardarkari baten atzean dagoenean, eta /jj/ gainerako kasuetan.

Hizlariek, baina, era askotara ebakitzen dute fonema hori. Esaterako: «joan» hitza [gj o a n] gisa ebakitzen dute zenbaitek, baina [x o a n] gisa beste zenbaitek. «j» duten hitz guztien ebakerak aztertu eta zuzendu dira (628 fitxategi).

- «tt» eta «dd» soinuak ere ondo aztertu dira, hizlari askok ez baitituzte hala ebakitzen, «t» eta «d» gisa baizik, hurrenez hurren (kasurik onenean).

- Transkribatzaileak «g» ortografiko guztiak /g/ bihurtzen ditu, baina hizlariek ez dute beti hala egiten. Esaterako: «teknologia» hitza [t e k n o l o g i a] gisa ebakitzen dute zenbaitek, baina [t e k n o l o x i a] gisa beste zenbaitek. Ez da gauza bera gertatzen «ge» eta «gi» kateak dituzten hitz guztiekin, eta, beraz, ez da erraza arazo horri irtenbide sistematiko bat aurkitzea. Letra-kate hori duten hitzen kopurua hain handia izanik (1.472 fitxategi), zenbait atzizki eta letra-kate hautatu dira («logia», esaterako) eta horiek baino ez dira zuzendu.

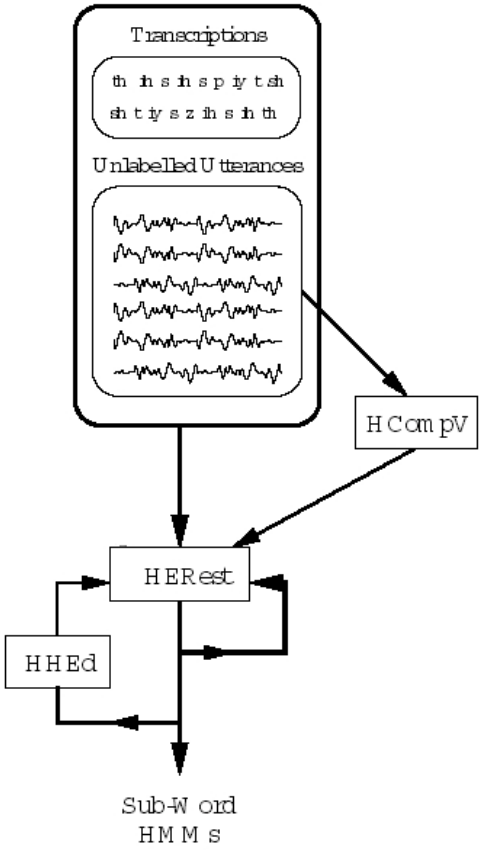
- Transkribatzaileak «i» osteko «l» eta «n»ak palatalizatzeko aukera ematen du, baina, hiztun guztiek ez dituzte fonemok bustitzen. Kopurua oso handia baita (3.052 fitxategi) eta nahiko testuinguru zehatzean gertatzen denez, ez dira zuzendu, izen berezien atalean izan ezean («Pili», esaterako).

4.2 Eredu akustikoak entrenatzea

Trifonema-eredu akustikoak sortzeko, lehendabizi fonemen HMMak sortu dira, gaussiar bakarrekoak. Lehendabizi, batezbesteko eta bariantza berdina duten HMMak sortu dira fonema bakoitzerako, prototipo batez baliatuta. Prototipo horren zeregina da ereduaren tipologia zehaztea, eta, hala, ezkerretik eskuinerako bost egoerako eredu bat hautatu da (lehen eta azken egoerak sarrera eta irteera dira), eta, seinale akustikoa parametrizatzeko, 39 MFCC (*Mel Frequency Cepstral Coefficients*) parametro baliatu dira batezbestekoak eta bariantzak kalkulatzeko: oinarrizko MFCC parametroak (13) gehi delta koefizienteak (13) gehi azelerazio-koefizienteak (13). Prototipoari buruzko xehetasun gehiago jakin nahi izanez gero, jo E.3 eranskinera.

HTKren HCompV tresnarekin, audio-fitxategi guztiak aztertu eta batezbesteko eta bariantza orokorrak konputatu dira ereduen hasierako balioak sortzeko. Energia-normalizaziorik ez da erabili, aukera hori ezin baita erabili denbora errealeko ezagutze-

prozesuak egiteko (energia normalizatzeko, fitxategi osoa irakurri behar da aurrez). Ondoren, HERest tresnarekin, batezbestekoak eta bariantzak entrenatu eta berrentrenatu dira (ikus 4.2 irudia).



4.2 irudia Fonemak HTK bidez entrenatzeko prozesua

Fonemen ereduak entrenatutakoan, trifonema edo fonema testuingurudunak sortu dira HLed eta HHEd tresnez baliatuz. Trifonemek zabaldu egiten dituzte fonemen mugak, eta, hala, kontuan hartzen dute fonemek aurrean eta atzean duten soinua: a-S+i trifonemak, esaterako, «a» baten ostean eta «i» baten aurrean dagoen «x» modelatzen du.

Hala eta guztiz ere, train azpicorpusa erabiliz sortu dira trifonemak, eta, beraz, litekeena da sortu eta modelatu ez diren trifonemak agertzea ezagutze-prozesuan. Hala izanez gero, HTK-k ezin izango luke ezagutze-prozesua burutu, eta bertan behera geldituko litzateke. Hori saihesteko, trifonema guztiak sortu dira automatikoki (28.831), eta antzeko trifonemekin lotu dira *cluster*retan multzokatuta. Antzeko trifonemak zein diren jakiteko, ezaugarrien araberako multzoak egin dira (ikus 4.1 taula) Urrutia eta

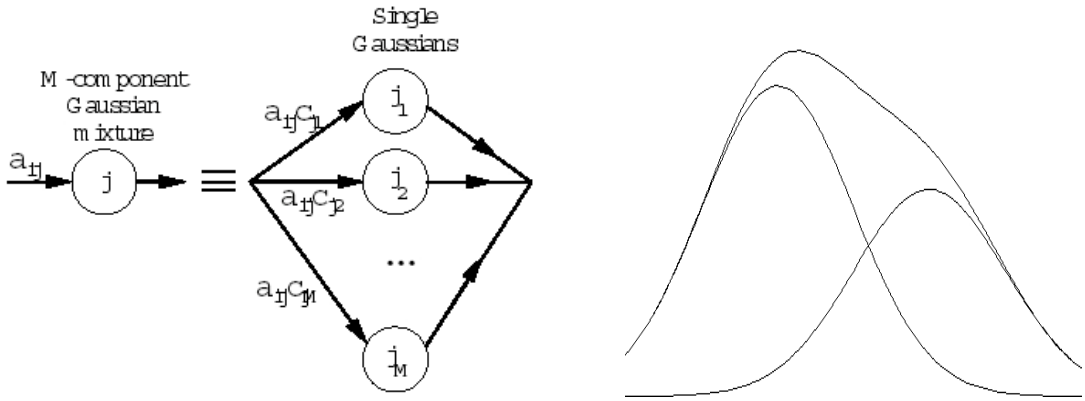
Etxebarriak egindako sailkapenaren arabera [17], eta erabaki-zuhaitz bidez klasifikatu dira:

```
EU_consonant p b t c d k g tS ts
ts` gj jj f B T D s s` S x G
Z m n J l L r rr
EU_sonora gj jj B D G Z m n J l
L r rr b d g
EU_sorda p t c k tS ts ts` f T s
s` S x
EU_plosive p b t c d k g
EU_affricate tS ts ts` gj
EU_fricative jj f B T D s s` S x
G Z
EU_nasal m n J
EU_liquid l L r rr
EU_lateral l L
EU_vibrant r rr
EU_bilabial p b B m
EU_dental t d D
EU_palatal c tS gj jj J L
EU_alveolar s` s n l r rr ts ts`
EU_velar k g x G
EU_prepalatal Z S
EU_apicoalveolar s ts
EU_postalveolar s` ts`
EU_plisiv_sorda p t c k
EU_plosiv_sonora g b d
EU_affric_sorda tS ts ts`
EU_fricat_sorda f T s s` S x
EU_fricat_sonora jj B D G Z
EU_bilab_sonora B b m
EU_dent_sonora d D
EU_palatal_sonora gj jj J L
EU_palatal_sorda c tS
EU_alveol_sonora n l r rr
EU_alveol_sorda s s` ts ts`
EU_velar_sonora G g
EU_velar_sorda k x
EU_plosiv_bilab p b
EU_plosiv_dent t d
EU_plosiv_velar k g
EU_affric_palatal tS gj
EU_affric_alveolar ts ts`
EU_fricat_alveolar s s`
EU_fricat_velar x G
EU_liq_alveolar l r rr
EU_vocal a e i o u
EU_frontal i e
EU_post o u
EU_cerrada i u
EU_semicerrada e o
EU_redonda o u
EU_no_redonda a e i
```

4.1 taula Euskararako zehaztutako talde fonetikoak, trifonemak multzokatzeko erabaki-zuhaitzetarako erabiliak

Trifonema-ereduak berriro berrentrenatu ondoren, irteerako probabilitatea modelatzen duten gaussiarren kopurua handitu da, modu progresiboan. Horretarako, HHEd tresna erabiltzen da, eta 2, 4, 8, 16, 32 eta 64 gaussiarreko ereduak sortu dira, bakoitzaren emaitzak ebaluatu ahal izateko.

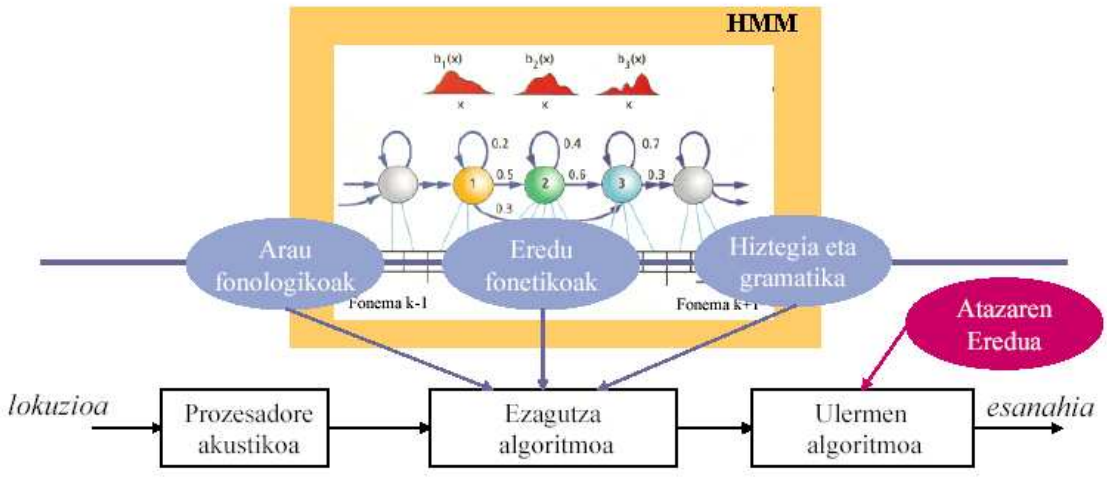
Dentsitate jarraituko irteera-banaketa anizkoitzak erabiliz, parametro jarraituak modelatzeko diseinatuta dago, batez ere, HTK; izn ere, irteerako probabilitate-dentsitatearen banaketa gaussiarren nahasturazko ereduak (GMM) erabiliz modelatzen da. 4.3 irudian, ageri da nahastura bat nola deskonposatzen den gaussiar bakunetan, baina GMMei buruzko xehetasun gehiago nahi izanez gero, jo *HMMren oinarriak* izeneko eranskinera.



4.3 irudia Gaussiarren nahastura

4.3 Testeko ezagutze-prozesua

Hizketa ezagutzeko prozesuaren eskema orokorra 4.4 irudian ageri da. Hizketa-seinalearen parametroak erauzi ondoren (lan honetan, 39 MFCC), ezagutza-algoritmoa aktibatzen da, Viterbiren algoritmoan oinarrituta. Ezagutza-algoritmoak eredu akustiko edo fonetikoak (HMMak) erabiltzeaz gainera, hiztegi bat eta gramatika bat ere behar du, sistemaren lexikoiko hitz bakoitzaren fonema-katea adierazteko eta hitzen segidak adierazteko, hurrenez hurren. Azkenik, ulermen-algoritmo bat aplikatzen da goragoko maila batean, ataza jakin bati loturik.



4.4 irudia Hizketa ezagutzeko prozesuaren eskema orokorra

Lan honetan, hitz isolatuak ezagutzeko sistema bat diseinatu denez, ez da behar izan, ez gramatikarik, ez ulermen-algoritmorik. Nahikoa izan da hiztegi bat sortzea eta, gramatika konfiguratzean, hitz solteak jasoko direla adieraztea.

Viterbiren algoritmoa HTK-ko HVite tresnarekin maneiatzen da.

4.4 Emaitzen analisia

Emaitzak analizatzeko, HTK-ko HResults tresna erabili da. Ezagutze-prozesuaren irteerak fitxategien transkripzio ortografikoekin alderatzen ditu tresna horrek, eta kontaketa sinple bat egiten du. Ehunekotan erakusten du emaitza.

5. PROBEN EMAITZAK

Hizketa ezagutzeko sistemaren testa egiteko, hiru azpicorpus erabili dira:

1. Test azpicorpuseko komandoak (105-316 sesioak): Azpicorpus horretan, hitz bakar bat duten fitxategiak soilik hautatu dira; hala, gramatika bat sortzea saihestu da. Guztira, 11.859 audio-fitxategi erabili dira. Hiztegiaren tamaina 435 hitzekoa da; horrek esan nahi du komando bakoitza 27,26 aldiz agertzen dela test azpicorpuseko 75 sesioetan eta, beraz, hitz horiek, nahiz eta beste hizlari batzuek esanda, badaude train azpicorpusean.

2. Bat-bateko hizketako izen bereziak eta leku-izen bereziak (013-015 sesioak): Azpicorpus horretan ere hitz bakar bat duten fitxategiak soilik hautatu dira. Sesio guztietako fitxategiak hartu dira, sesio horiek ez baitira erabili sistema entrenatzeko (hizketa irakurria besterik ez, 025-306 fitxategiak). Guztira, 584 fitxategi dira, eta hiztegiaren tamaina 206 hitzekoa da. Hitz hauek, oro har, ez daude train azpicorpusean.

3. Bat-bateko hizketako izen bereziak eta leku-izen bereziak (2. ataleko azpicorpusa), ZientziaNeteko hiztegiaz. Anhitz proiektuko eszenatoki birtualean erabiliko den lexikoia kontuan izanda, hiztegi berri bat sortu da, baina 2. ataleko hizketa-seinaleak erabili dira. Hortaz, guztira 584 fitxategi dira, baina hiztegiaren tamaina 19.383 hitzekoa da.

5.1 taulan, azpicorpus bakoitzaren xehetasunen laburpen bat ageri da.

Testa	Azpicorpusa	Hizketa mota	Fitxategi kopurua	Hiztegiaren tamaina
1.a	Test azpicorpuseko komandoak	Irakurria	11.859	435
2.a	Corpus osoko izen eta leku-izen bereziak	Bat-batekoa	584	206
3.a	Corpus osoko izen eta leku-izen bereziak	Bat-batekoa	584	19.383

5.1 taula Test-proben xehetasunen laburpena

Aipatu beharra dago 3. probarako hiztegia egiteko, ZientziaNeteko lexikoia eta Elhuyar hiztegia erabili direla. Izan ere, ZientziaNeteko hiztegia «garbitu» gabe dago; hau da, erabiltzaileek sartutako hitz guztiak daude han, nahi gabe gaizki idatzitakoak, akats ortografikodunak, erdarazko hitzak eta jolasean sartutako hitz zentzugabeak barne.

Horrenbestez, ZientziaNet garbitzeko, Elhuyar hiztegiaren sarrerak erabili dira (60.502 sarrera), eta han agertzen diren ZientziaNeteko hitzak baino ez dira hautatu. Gainera, artikulu singularra eta plurala eta zenbait posposizio (genitibo singularra eta plurala eta leku-genitiboa) ere kontuan izan dira, ZientziaNeteko hitz asko ez baitaude hitz-erro gisa idatziak. Hala, jatorrizko 52.761 hitzeko lexikoitik 19.383 sarrerako hiztegi bat sortu da. Prozesu horretan, izen berezi asko galdu dira, eta, hortaz, komeni da, etorkizuneko lanetarako, erauzketa hori xehetasun gehiagoz egitea, baina lan honetarako nahikoa dela uste dugu.

Test-proben emaitzak 5.2 taulan jaso dira, okerreko hitzen ehunekotan, ezagutze-prozesuan erabilitako gaussiar kopuruaren arabera.

Testa	Gaussiar kopurua						
	1	2	4	8	16	32	64
1.a	% 4,98	% 3,04	% 1,45	% 1,30	**% 1,15**	% 1,32	% 1,96
2.a	% 10,10	% 8,05	**% 5,72**	% 6,16	% 5,99	% 6,34	% 6,34
3.a	% 24,83	% 20,21	% 13,36	% 12,33	**% 10,96**	% 12,84	% 18,66

5.2 taula Test-proben emaitzak (WER), gaussiar kopuruaren arabera

6. ONDORIOAK

Emaitzarik onenak 16 gaussiarreko probetan lortu dira (2. proban, lau gaussiarreko probaren emaitzetik oso hurbil dago 16koa). 16 gaussiar baino gehiago erabilita, emaitzak kaxkartu egiten dira pittin bat, beharbada HMMak hizlarien ezaugarritara gehiegi moldatzen direlako (gainadaptazioa).

Bestalde, emaitzak ikusita, nabaria da bi faktore nagusik eragiten diotela ezagutze-prozesuari: batetik, ezagutu beharreko hitza entrenatzeko corpusean agertzen den ala ez. Hitzak entrenatuta baldin badaude, , beste hizlari batzuen bidez izanik ere, askoz hobeto ezagutzen ditu sistemak. Bestetik, hiztegiaren tamainak eragin argia du; hiztegia zenbat eta handiagoa, orduan eta nahaste handiagoa sortzen du sisteman, eta beste teknika batzuk erabili behar dira hitzen (HMM-kateen) probabilitateak doitzeko: hitz konektatuetan, esaterako, gramatika bat zehaztea, edo, hizketa jarraituan, hizkuntza-modelatzea. Bestalde, hitzak stema eta hondarkitan zatitzeko aukera ere aztertu behar da [17].

AnHitz proiektuko eszenatokirako ezagutze-sistema diseinatzeko, bi aldaketa proposatzen dira, emaitzak ikusita:

- Hiztegia mugatzea eta gramatika bat sortzea: Proba gehiago egin beharko dira hiztegi berarekin baina galdera motako gramatika batekin hiztegia handitzeak sortzen duen nahasmena nola portatzen den ikusteko. Hala ere, ZientziaNeteko hiztegia hobeto aztertu behar da, eta selektiboagoa izan behar da; balioetsi behar da ea merezi duen behin bakarrik agertzen diren hitzak (hiztegiaren % 50) erabiltzea.

- HMM ereduak sortzeko, corpus osoa erabiltzea: lan honetan, 155 sesio erabili dira, eta gainerako 75ak testa egiteko erabili dira. Hortaz, sistema nabarmen hobetu daiteke entrenatzean 75 sesio horiek kontuan hartzen badira.

7. AIPAMENAK

[1] Huang, Xuedong; Acero, Alex; Hon, Hsiao-Wuen (2001): *Spoken language processing: a guide to theory, algorithm, and system development*. Prentice Hall PTR, New Jersey.

[2] Deng, Li & Huang, Xuedong (2004): *Challenges in adopting Speech Recognition*, Communications of the ACM, Vol. 47, No 1, New York.

[3] DARPA's EARS Kickoff Meeting, 2002ko maiatzaren 9-10, Viena.

[4] DARPA's EARS Conference, 2003ko maiatzaren 21-22, Boston.

[5] Furui, Sadaoki (2002): *Recent progress in spontaneous speech recognition and understanding*. Proc. IEEE Workshop on Multimedia Signal Processing, Dec. 2002.

[6] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens (2007): *Automatic speech recognition and speech variability: A review*. Speech Communication 49, pp 763-786.

[7] Stanley F. Chen, Brian Kingsbury, Lidia Mangu, Daniel Povey, George Saon, Hagen Soltau, and Geoffrey Zweig (2006): Advances in speech transcription at IBM under the DARPA EAR program. IEEE, Transactions on audio, speech and language processing, Vol. 14, No 5.

[8] Hernaez, I., Luengo, I., Navas, E., Zubizarreta, M., Gaminde, I., Sanchez, J. (2003): *The basque speech_dat (II) database: a description and first test recognition results*, In Eurospeech-2003, 1549-1552.

[9] Hernaez, I., Navas, E., Sánchez, E., Madariaga, I., Gaminde, I., Zalbide, X. (2002): *BIZKAIFON: A sound archive of dialectal varieties of spoken Basque*, EHUko AhoLab taldea, Bilbo.

[10] López de Ipiña, K., Torres, I., Oñederra, L. (1995): *Design of a Phonetic Corpus for a Speech Database in Basque Language*, in Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Technology. Madrid, Spain, 18-21 September, 1995. Vol 1, pp. 851-854.

[11] López de Ipiña, K., Torres, I., Oñederra, L. (1998): *A speech database in Basque language*, Workshop on Language Resources for European Minority Languages, May 27 1998, Granada, Spain.

[12] López de Ipiña, Karmele (2003): *Euskarazko hizketa jarraituaren ezagutza, metodo estokastikoen bidez*. Doktorego-tesia, EHU, Donostia.

[13] Odriozola, Igor (2007): *Bizkaierazko hitz isolatuak ezagutzeko sistema baten garapena*. Karrera Amaierako Proiektua, AhoLab laborategia, EHUko Elektronika eta Telekomunikazioak Saila, Bilbo.

[14] Xabier Zalbide, Iñaki Gaminde, Inma Hernaez, Maren Zubizarreta, Eva Navas (2003): *Euskararko SAMPA kodeaz*, AhoLab laborategia, EHUko Elektronika eta Telekomunikazioak Saila, Bilbo.

[15] Young, S.; Odell, J.; Ollason, D.; Valchev, V.; Woodland, P.: *The HTK book*, HTKren eskuliburua, Cambridgeko Unibertsitatea, 2000ko ekaina.

[16] HTKren web orri ofiziala: http://htk.eng.cam.ac.uk/

[17] Rotovnik, T., Maucec, M.S., Kacix. Z. (2007): *Large vocabulary continuous speech recognition of an inflected language using stems and endings*. Speech Communication, Vol.49, No.6, pp.437-452

ERANSKINAK

E.1 ZientziaNet lexikoiko lagina

12986 energia
4866 eta
4735 ur
3530 da
3165 zer
3077 eguzki
2916 animali
2788 euskal
2545 historia
2450 mota
2450 kutsadura
2428 petrolio
2244 herri
2232 zelula
2069 berriztagarri
2045 elikadura
2034 nuklear
2011 gaixotasun
1804 sistema
1782 lur
1780 zentral
1772 sumendi
1753 klima
1708 efektu
1690 zientzia
1634 ozono
1603 paper
1550 elektriko
1508 aparatu
1495 kirol
1484 egipto
1470 ikatz
1461 ziklo
1395 planeta
1379 bale
1358 azido
1330 aldaketa
1287 geruza
1265 baso
1243 ekosistema
1235 euri
1234 mundu
1219 ugalketa
1217 hies
1217 eoliko
1201 eboluzio
1196 berri
1181 mineral
1170 berotegi
1167 elektrizitate
1147 hartz
1146 droga
1127 gizaki
1124 giza
1123 klonazio
1117 biografia
1098 tabako
1096 telefono
1080 natural
1077 galileo
1074 lurrikara
1069 nola
1068 termiko
1059 ilargi
1050 basamortu
1049 marrazo
1047 unibertso
1040 dira
1027 zarata
996 hidrauliko
989 elikagai
987 anorexia
985 pila
951 osasun
947 zaldi
945 transgeniko
943 ama
940 minbizi
939 atmosfera
897 definizio
888 telebista
862 teknologia
851 sorrera
851 gas
845 ezaugarri
831 arazo
828 edison
817 zuhaitzak
812 diabete
806 beltz
805 informazio
804 mapa
786 gripe
778 musika
776 itsas
757 ez
738 teoria
736 zuri
713 dieta
713 biomasa
706 de
697 newton
677 pluton
676 zura
666 egur
663 urre
634 alkohol
630 zirkulazio
629 bizitza
624 zuntz
622 gerra
606 darwin
593 galilei
560 genetika
557 zuria
513 erdi
508 negutegi
471 argia
467 big
463 makina
462 ibaiak
461 txakurra
461 hizkuntza
461 bang
453 ibaia
452 zulo
452 genoma
450 thomas
450 kolera
446 zuhaitza
446 nekazaritza
443 internet
439 franklin
436 landareak
431 arnas
427 kontsumoa
427 itsasoa
423 hondakinak
422 argazkiak
421 landare
420 zientifikoa
420 tigrea
417 otsoa
412 zertarako
411 garapen
411 bihotza
409 sabana
409 lehoia
405 ondorioak
405 jatorria
401 gorputza
399 material
397 einstein
394 proteinak
393 iturriak
390 saturno
387 kimika
386 begia
383 arrantza
382 altzairua
377 prozesua
376 materia
376 lehen
376 birziklapena
375 tundra
374 san
373 nerbio
373 klimatiko
371 erregai
371 curie
370 euskadi
368 benjamin
365 industria
362 izarrak
360 historiaurre
355 euskara
355 burdina
352 zaborra
352 bulimia
351 karbono
350 genetikoa
347 y
347 marea
347 arkimedes
346 eragina
344 tropikala
342 lan
342 jupiter
342 funtzionamendua
342 aroa
338 egitura
337 katua
337 bizi
336 zuen
330 odol
330 fosilak
329 izurdea
328 arranoa
327 landareen
326 taula
326 plastikoa
324 fotosintesia
324 ekonomia
324 asma
324 adn
323 bat
322 obesitatea
322 istripuak
321 odola
320 zuloa
320 grezia
319 kimikoa
319 fauna
318 medikuntza
318 gaur
313 leonardo
310 zubia
310 2
308 geologia
306 ekologia
306 artea
304 kultura
304 geotermikoa
302 nagusia
302 du
301 garapena
300 europako
299 optikoa
298 homo
297 kimikoak
297 informatika
297 bioteknologia
295 egin
294 vinci
293 protokoloa
292 menpekotasuna
292 liseri
291 irratia
290 haizea
290 emakumea
290 atalak
289 europa
289 eklipsea
289 armiarma
...

...	1 abioia	1 abandonatua	1 20050
1 acicularis	1 abioen	1 abanataila	1 1996
1 achwann	1 abioa	1 abaluazioa	1 1986
1 acerobacter	1 abintalak	1 abaltzantza	1 1985
1 acerifolia	1 abiles	1 abalone	1 1976
1 acdc	1 abietazeoak	1 abalon	1 1972
1 accsi	1 abiazioaen	1 abalantxak	1 1963
1 accidente	1 abiazioa	1 abal	1 1953
1 acb	1 abiarr	1 abako	1 1950
1 acartia	1 abiardura	1 abakantoa	1 1947
1 abzisikoa	1 abhhhkh	1 abaitua	1 1945
1 abzisiko	1 abeztia	1 abair	1 194
1 abusoak	1 abetoa	1 abaia	1 1936
1 abusimbel	1 abestruz	1 abadiñoko	1 1935
1 aburuz	1 abeslarien	1 abadiak	1 1934
1 abulon	1 abesbetzak	1 aba	1 1931
1 abuloia	1 abesbatzak	1 aazoa	1 192
1 abujetak	1 abesa	1 aautomatizazioa	1 1915
1 abujero	1 abertzaleen	1 aat	1 1911
1 abuferal	1 abertzaleak	1 aasiera	1 1885
1 abtentzioa	1 abertzalea	1 aaser	1 1881
1 abstinentzi	1 abertzale	1 aaron	1 1873
1 abstenia	1 abertsaletasun	1 aariskua	1 1870
1 abstakzioa	1 aberriko	1 aardvarks	1 1861
1 absolutoa	1 aberri	1 aanson	1 1858
1 absolutismo	1 aberats	1 aae	1 1857
1 absoltua	1 abera	1 99	1 1852
1 absentismoa	1 abenturazko	1 96	1 1847
1 absentismo	1 abenturako	1 9317	1 1837
1 abrazioa	1 abentano	1 9230	1 1836
1 abrasioa	1 abenduko	1 91	1 1835
1 abrasio	1 abeltzautza	1 8889	1 1834
1 abraitza	1 abeltzantzako	1 84568246	1 1818
1 abrahan	1 abeltzaitza	1 84001	1 1814
1 abra	1 abeltzaintzari	1 84	1 1811
1 abotua	1 abeltzainen	1 77777	1 1804
1 abotu	1 abeltzai	1 776	1 180
1 abotoa	1 abeltza	1 7175	1 1783
1 aboto	1 abelttzaintza	1 7128	1 1781
1 abortoaren	1 abeltaizta	1 696	1 1780
1 abortastu	1 abeilanaketa	1 684	1 1740
1 aboroa	1 abegitsu	1 65465456	1 1727
1 aborijenoa	1 abedul	1 60	1 1664
1 aborijeno	1 abea	1 581561	1 1644
1 aborijenak	1 abe	1 56	1 1642
1 aborijena	1 abdoulaye	1 541736	1 1550
1 aborigenoa	1 abdominalak	1 52	1 155
1 aborigenes	1 abdominala	1 49423	1 1512
1 aborigenea	1 abdominaak	1 4865	1 1511
1 aboridene	1 abdomena	1 47	1 151
1 aboribene	1 abdelaziz	1 436	1 1500
1 aboregenak	1 abdea	1 4343	1 148
1 abono	1 abbadya	1 4239	1 1470
1 abonatzeko	1 abash	1 41	1 1453
1 aboluzioa	1 abarra	1 407	1 1422
1 abogadroren	1 abaro	1 406	1 137
1 abogadro	1 abarkak	1 338	1 1348
1 abodi	1 abar	1 330	1 134
1 abnimaliak	1 abantala	1 33	1 121
1 abizenen	1 abantailk	1 2436	1 11631
1 abitominosia	1 abantailaketa	1 224	1 112
1 abitominosi	1 abantail	1 2206	1 1080
1 abitat	1 abantaia	1 219	1 102
1 abitaminosis	1 abantai	1 210	1 101
1 abitamina	1 abanta	1 2036	1 10081
1 abispa	1 abaniko	1 2024	1 1001
1 abionika	1 abandonoa	1 2009	1 1000
1 abioien	1 abandonatzea	1 2008	1 08

E.2 ASR-3200 datu-baseko datu-fitxategien lagina

```
[System]
Script=121
Version=0121_1
Board=2;NIDAQ
Coding=PCM;Linear
Frequency=16000
Channels=2
ByteFormat=01
Delimiter=>-<
CharacterSet=ANSI
Ansi codepage=1252
Dos codepage=437
Male/female differentiation=0
Memo=Basque data collection. 1 Shure SM10A close talk microphone, 1 Shure SM58. One speaker
and one facilitator operating the recording equipment, room 10 Softec-Basauri, 45 square
meters,far wall (1)

[Channels]
1=SHURE SM10A>-<CLOSE_TALK
2=SHURE SM58>-<DESKTOP
3=>-<
4=>-<

[Versions]
DSDR=Version 7.2.2
DSVS=Version 6.41

[Info states]
1=Name>-<Leire Basarte Zuberogoitia>-<
2=Speaker>-<10200>-<
3=Age>-<29>-<
4=Sex>-<Female>-<
5=Region of birth>-<Batua Bizkaia>-<
6=Region of youth>-<Batua Bizkaia>-<
7=Language_Level>-<High level non native>-<
8=Environment>-<Office>-<
9=Location Reference>-<Softec Basauri 10>-<
10=EXN>-<David Santamaria Vazquez>-<
11=Position>-<Far Wall_01>-<
12=Audio>-<Off>-<
13=Size>-<30+ m²>-<
14=Remarks>-<>-<
15=Sheet_ID>-<1>-<

[Session]
Number of recordings=316
Sheet number=1
Directory=c:\adb_0121\data\scr0121\26\01212601\r1210001
Imported sheet file=C:\BASQUE\Script\bae121_1.psh
RecDate=07 jun 2004
RecTime=18:04:58
Record duration=112'0
Record session=1

[Record states]
1=2>-<>-<Mesedez, egon ixilik. Inguru-zarata grabatzen ari gara.>-<1024>-<1281024>-
<U0001001.WAV>-<>-<-1>-<-1>-<SY1>-<0>-<>-<
2=1>-<>-<Gonbidatu norbait zure urtebetetze egunean antolatu duzun ospakizunera erantzungailu
automatikoan mezu bat utziz.>-<1024>-<4852224>-<U0001002.WAV>-<>-<-1>-<-1>-<SS1>-<0>-<>-<
3=1>-<>-<Bidaia-agentzia batera deitu bidaia bat erreserbatzeko. Galdetu irtetzeko eta
heltzeko orduak zein diren, prezioa, hotel mota eta kokalekua, ikuskizunak, ordaintzeko erak,
etab.>-<1024>-<5748224>-<U0001003.WAV>-<>-<-1>-<-1>-<SS2>-<0>-<>-<
```

4=1>-<>-<Banketxera deitu eta zure kontuaren saldoa zein den galdetu; gero, beste kontu batera diru-transferentzia bat egin, eta bukatzeko, burtsaren mugimenduei buruz galdetu (akzioen prezioak, salmentak, erosketak, ...).>-<1024>-<5236224>-<U0001004.WAV>-<>-<-1>-<-1>-<SS3>-<0>-<>-<
5=1>-<>-<Zinema edo antzoki batera deitu eta egiaztatu pelikula edo antzezlan jakin baterako eserleku librerik dagoen.>-<1024>-<5428224>-<U0001005.WAV>-<>-<-1>-<-1>-<SS4>-<0>-<>-<
6=1>-<>-<Liburu edo musika denda batera deitu eta galdetu ea liburu edo CD zehatz bat duten; bidenabar, salneurria eta argitaraldia ere galdetu.>-<1024>-<5620224>-<U0001006.WAV>-<>-<-1>-<-1>-<SS5>-<0>-<>-<
7=1>-<>-<Gutun komertzial labur bat diktatu gai hauetako bat aukeratuz: eskaera bat, negozio bidaia bat, hitzaldi bat, edo negoziaketak.>-<1024>-<5018624>-<U0001007.WAV>-<>-<-1>-<-1>-<SS6>-<0>-<>-<
8=1>-<>-<Zein da zure aitaren jaiotza-eguna?>-<1024>-<257024>-<U0001008.WAV>-<>-<-1>-<-1>-<T1>-<0>-<>-<
9=1>-<>-<Zein da gaurko data?>-<1024>-<180224>-<U0001009.WAV>-<>-<-1>-<-1>-<T2>-<0>-<>-<
10=1>-<>-<Noiz hasiko dira zure hurrengo oporrak?>-<1024>-<180224>-<U0001010.WAV>-<>-<-1>-<-1>-<T3>-<0>-<>-<
11=1>-<>-<Zer ordu da?>-<1024>-<141824>-<U0001011.WAV>-<>-<-1>-<-1>-<T4>-<0>-<>-<
12=1>-<>-<Zer ordutan jan duzu azkenekoz?>-<1024>-<116224>-<U0001012.WAV>-<>-<-1>-<-1>-<T5>-<0>-<>-<
13=1>-<>-<Zein hiri edo herritan egiten duzu lan?>-<1024>-<129024>-<U0001013.WAV>-<>-<-1>-<-1>-<IWn1>-<0>-<>-<
14=1>-<>-<Zein hiri edo herritan pasa zenuen zure haurtzaroa?>-<1024>-<116224>-<U0001014.WAV>-<>-<-1>-<-1>-<IWn2>-<0>-<>-<
15=1>-<>-<Zein da zure ama edo aitaren izena?>-<1024>-<103424>-<U0001015.WAV>-<>-<-1>-<-1>-<IWn3>-<0>-<>-<
16=1>-<>-<Gehien gustatzen zaizun aktorearen izena eman.>-<1024>-<141824>-<U0001016.WAV>-<>-<-1>-<-1>-<IWn4>-<0>-<>-<
17=1>-<>-<Enpresa baten izena eman.>-<1024>-<141824>-<U0001017.WAV>-<>-<-1>-<-1>-<IWn5>-<0>-<>-<
18=1>-<>-<Irakur ezazu zure abizena letraz letra.>-<1024>-<295424>-<U0001018.WAV>-<>-<-1>-<-1>-<CA1>-<0>-<>-<
19=1>-<>-<Euskeraz hitz egiten al duzu?>-<1024>-<154624>-<U0001019.WAV>-<>-<-1>-<-1>-<IWa1>-<0>-<>-<
20=1>-<>-<Kantoneraz hitz egiten al duzu?>-<1024>-<167424>-<U0001020.WAV>-<>-<-1>-<-1>-<IWa2>-<0>-<>-<
21=1>-<>-<Gehien menperatzen duzun hizkuntzaren izena esan.>-<1024>-<193024>-<U0001021.WAV>-<>-<-1>-<-1>-<IWn6>-<0>-<>-<
22=1>-<>-<Ezagutzen duzun pertsona baten telefono zenbakia esan.>-<1024>-<205824>-<U0001022.WAV>-<>-<-1>-<-1>-<FN1>-<0>-<>-<
23=1>-<>-<Ezagutzen duzun beste pertsona baten telefono zenbakia esan.>-<1024>-<308224>-<U0001023.WAV>-<>-<-1>-<-1>-<FN2>-<0>-<>-<
24=1>-<>-<Ezagutzen duzun beste pertsona ezberdin baten telefono zenbakia esan.>-<1024>-<282624>-<U0001024.WAV>-<>-<-1>-<-1>-<FN3>-<0>-<>-<
25=2>-<>-<Zaragozako unibertsitateko irakaslea ere bazara.>-<1024>-<282624>-<U0001025.WAV>-<>-<-1>-<-1>-<ISp1>-<0>-<>-<
...

[Validation states]
1=>-<-1>-<-1>-<QUA:X>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001001.WAV>-<1024>-<1281024>-<>-<>-<>-<
2=aupa Zuber aber badakizu datorren hamalauan nire urtebetetzea dela ezta bueno ba astelehenean jausten da baina bueno ez dakit asteburu horretan hamabi hamahiru uste dut dela ala hemeretzian eta hogeian zeozer egingo dugun badakizu ere Kristinaren urtebetetzea ondino ez {FRA} dagoela ospatuta eta klaro berarekin egitea ba igual hoberena izango da ez dakit {FIL} aita badakizu daukala txoko bat eta ez dakit bion artean egingo dugun han txokoan {FRA} ala bueno ba ez dakit ondo zelan egingo dugun baina gauza da hemeretzi gauean ezin dugula ba nik {FIL} Enekoren lagunaren partez beste urtebetetze bat daukadalako eta ba noizbait egin behar dugu baina klaro denok konforme egoteko ba ez dut oso ondo ikusten aber zelan ikusten duzun zuk zuk hitzegiozu gure baduzu Ainhoa eta Kristinagaz baita eta Veronica Rafa eta hauegaz be hitz {FRA} egingo dugu por que klaro denok batera egin beharko dugu eta nik nahiko nuke ia denok {FIL} geratzen {FRA} geratzen garen baina ba zail samar ikusten dut eta ja ba klaro ekaina da batzuk {FIL} badoazte kanpora eta bueno ba aber zela egiten dugun bale deitu mesedez ba aber {FIL} ez dakit Kristinagaz hitz egiten duzunean edo {FRA} eta zeozertan geratuko gara bale benga ba ederto agur>-<-1>-<-1>-<QUA:B>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001002.WAV>-<21320>-<4826848>-<>-<>-<>-<
3=kaixo {FIL} begira ba aber {FIL} gura dut {FIL} bidai bat egitea Egiptoruntza gauza da ba {UNI} ba ez dudala tenporada altuan {FRA} egon {FIL} joan nahi ba {UNI} nahiko diferentzi

dagoela prezioetan eta holan eta {UNI} nuke ba horrelako oferta on bat hartzea baina {FIL} ez dut {FIL} bidai holan ahal ba da ba bidaia pixka bat luzeagoa izatea ba ez ez diot hamabost egunetara heltzea ezta {FRA} baina bederatzi egunetan ba nahiko ondo egongo litzateke gero ba prezioak eta jakiteko gutxi gora {FRA} behera {FRA} nola nola doazten eta hotelak be gero badakit ba trabesia bat dagoela Nilotik hotel {FRA} bueno hotel flotantetan edo holan izango dela ezta {FIL} mota desberdinak {FRA} badela badela badakit ba batzutan ba entzunda daukat ba nahiko jende ba gaixotu egiten dela ba janariengatik ba nahiko hotel ona gura dut izatea ba hori ba ebitatzeko ahal ba da eta {FRA} gero ba ez dakit ikuskizunak egongo direzen badakit Abus Imbelen be badagoela ikuskizun bat argiena eta holan eta ba ikusi nahiko nuke {FIL} ez dakit ordaintzeko erak zeintzuk izango diren {FIL} esango didazu ba ez dakit {FIL} igual orain {FIL} atal bat {FRA} edo hilabete barru beste atal bat edo zelan ordainduko dudan ba ez dakit ondo erreserbatu {FRA} egin egitea gustatuko litzaidake bi pertsonentzako eta bueno ba deituko zaituztet geroago ba eta esateko zelan diren gauzak eta holan ba hobeto hitz egiteko zuekin benga ba agur>-<-1>-<-1>-<QUA:C>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001003.WAV>-<19056>-<5585632>-<>-<>-<>-<
4=arratsaldeon {FIL} bai begira nire kontua hiru sei {FRA} zero zero zero zero zero da gura neuke jakitea ba zenbat {FIL} diru daukadan kontu horretan azken finean datorren hamalauean nire urtebetetzea denez badakit ba hogeita hamar urte egiten ditudanez ba kontu {FRA} gaztea edo dena delakoa ba itxi egingo didazuela eta badakit ba hori ba aldatu beharko dudala dirua beste kontu batetara {FRA} baina {FRA} zein zenbateko {FRA} interesa emango didazuen ba ez dakit eta gura neuke jakitea gutxi gora behera ze {FRA} motatako ba kontuak dauden ba {FIL} zein den hobeena zein den jakiteko eta horretara ba mugitzeko nire diru guztia ezta gero transferentzia bat egin nahi dut beste kontu batera senarrarena {FIL} bere kontua bost sei sei sei zero zero zero dugu eta horretara pasa gure nuke {FRA} ba gutxi gora behera ehun euro gero ba {FIL} akzioak ditut bai Telefonican eta orain {FRA} zer zenbateko prezioan dauden jakin gure nuke {FRA} ba saltzeko gogoak ditudalako hamasei eurotara heltzen denean bakoitza {FRA} ba {FIL} ba zuei ematea {FRA} ba aukera ba hori ba saltzeko zeuek {FIL} jakin dezazuen ba permisu edo nire permisurik gabe edo dena delakoa bueno ba deituko zaituztet beste momentu baten eta ba egongo gara>-<-1>-<-1>-<QUA:B>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001004.WAV>-<22928>-<5004404>-<>-<>-<>-<
5=aber ba bai {FIL} begira ba gure nuke Troya pelikula ikustea gaur gauean hamarretako {FRA} sesioarentzako ez dakit hamarrak {FRA} hamarrak eta laurden edo hamarrak eta laurdenetan egongo den baten bat ba gutxi gora behera ordu horietatik gero ba bi pertsona izango gara {FIL} gura nuke ba gutxi gora behera ba erditik {FIL} atzerantzako eserleku bat hartzea bueno bi kasu honetan eta ahal bada behintzat eta izkinetan baldin bada nik ez dut gure eta ba ez dakit {FIL} adibidez gela {FRA} hamasei sala hamasei ilarakoa baldin bada ba hamalautik gutxi gora behera egotea nahiko nuke hamabitik {FRA} beherakotzat beherakora baldin bada ba ba ez dut nahi eta orduan {FRA} ba esan mesedez ba ze motatako eserlekurik dagoen libre ba {FRA} hori hartzea ez dudalako nahi gero cinesa card daukadanez ba ia telefonoz {FRA} eskura erreserba egiten baldin badut ia cinesa horrek balioko didan jakin gura nuke badakizu ba hori ba bi entradak hartzerakoan puntuak hartzen ditudala eta ba {FRA} beste ez dakit sarrera bat beste egun baterako libre izateko edo dohainik {FIL} jakingo {FRA} nahi dut hori ia erreserba egiten badut aukera hori daukadan ala ez bueno ba listo>-<-1>-<-1>-<QUA:C>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001005.WAV>-<23752>-<5194580>-<>-<>-<>-<
6={FIL} epa begira ba zera galdera bat daukat CD bueno CD ez bideo baten {FIL} ba begira nago ia aurkitzen dudan ez dakit deskatalogatuta egon den Stray Catsena da oraintxe elkartu behar direz gainera hogei urte pasata baina hau aurreko denborena da {FRA} ez dakit ondo zelako izena daukan bez ez baina ba dakit Japonian {FIL} ba grabatu egin zela edo denadelakoa esandakoa ez dakit bere izenik {FIL} ezta noiz izan zen hogei urte izango direz igual {FIL} argitaratu {FRA} zena Europan badakit {FRA} deskatalogatuta dagoela eta ba ez dakit {FIL} Japonian edo Amerikan lortu daitekeen {FIL} jakin gura dut ia zeuek {FIL} lortu dezakezuen {FIL} bideo hori {FRA} ba ia lan egiten duzuen Alemaniarekin oi Alemaniarekin esango dut {FIL} Japoniarekin ala Amerikarekin {FRA} ba ba hori ba eskuratzeko mesedez ipini {FRA} nigaz kontaktuan nere telefonoa bederatzi lau lau lau lau lau bost da eta ba hori gura neuke jakitea ahal ba da {FRA} ba ipini nigaz mesedez kontaktuan ba nahiko {FRA} interesa daukadalako opari batentzako delako aspaldi nago hori bilatzen {FRA} eta nahiko interesatuta nago ez dakit diru aldetik zenbat izango dan {FRA} horri horri buruz ere hitz egin gura nuke {FIL} oso garestia ba da ba noski ez dut hartuko eta baina bueno {FIL} nahiko diru ordaintzeko prest nago mesedez deitu ondo>-<-1>-<-1>-<QUA:C>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001006.WAV>-<30412>-<5592576>-<>-<>-<>-<
7=aber Maria mesedez hartzazu bolia {FRA} eta eta folio batzuk orain {UNI} zaitudalako diktatuko dizudalako barkatu gutun komertzial bat bai {FIL} aber ez dakit zelan hasiko garen {FRA} ba adibidez {FRA} ba {FIL} probeedore jaun horri edo holako zeozer gure duzun moduan idatzi zeozer holan {FRA} sarreran ondo egoteko bueno {FIL} telefonoz hitz egin genuen moduan ba {FIL} ba hori idatzita nahi zenuenez ba hona ba gure {FRA} eskaera gure eskaera lehenengotan ehun bat boligrafo urdin gorri berde eta beltz beste ehun bat {FIL} atzingi {FIL} ehun bat zorroskilo {FRA} eta bostehun {FIL} orritako folio pakete {FRA} ba hogei bat

holan jarraitu ba ez dakit {FRA} betiko {FIL} eskaera eginez mesedez ba dakizu hau formala bakarrik da formaltzeko delako eta ba {FRA} bukatzeko gutunarekin ba ez dakit {FIL} esan {FRA} ba bueno zuk ikusten duzun moduan mesedez formala izatea eta jasta ederto bukatu dugu>-<-1>-<-1>-<QUA:C>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001007.WAV>-<32512>-<4901360>-<>-<>-<>-<
8=mila bederatziehun eta berrogeita seiko urtarrilak   bia>-<-1>-<-1>-<QUA:B>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001008.WAV>-<31880>-<206956>-<>-<>-<>-<
9=ekainak zazpi>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001009.WAV>-<61856>-<140988>-<>-<>-<>-<
10=uztailak hamasei>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001010.WAV>-<57228>-<141868>-<>-<>-<>-<
11=bostak bost gutxi>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001011.WAV>-<29944>-<111688>-<>-<>-<>-<
12=ordu bietan>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001012.WAV>-<34888>-<102192>-<>-<>-<>-<
13=Basaurin>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001013.WAV>-<42744>-<106352>-<>-<>-<>-<
14=Getxon>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001014.WAV>-<30920>-<86040>-<>-<>-<>-<
15=Javi>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001015.WAV>-<16640>-<68532>-<>-<>-<>-<
16=George Clooney>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001016.WAV>-<44320>-<110304>-<>-<>-<>-<
17=Softec>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001017.WAV>-<44836>-<108916>-<>-<>-<>-<
18=be A ese A erre te E>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001018.WAV>-<42664>-<247624>-<>-<>-<>-<
19=bai>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001019.WAV>-<28228>-<72060>-<>-<>-<>-<
20=ez>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001020.WAV>-<45232>-<98244>-<>-<>-<>-<
21=gaztelania>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001021.WAV>-<56520>-<126784>-<>-<>-<>-<
22=bederatzi lau lau hirurogeita sei hamahiru hamabi>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001022.WAV>-<28480>-<180884>-<>-<>-<>-<
23=bederatzi lau sei zero bi   hamasei {UNI} zazpi>-<-1>-<-1>-<QUA:C>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001023.WAV>-<26716>-<215268>-<>-<>-<>-<
24=sei zero zazpi berrogeita zazpi berrogeita sei berrogeita zazpi>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001024.WAV>-<34272>-<225472>-<>-<>-<>-<
25=Zaragozako unibertsitateko irakaslea ere bazara>-<-1>-<-1>-<QUA:A>-<NOI:0>-<SND:0>-<SPC:0>-<UTT:0>-<DST:0>-<U0001025.WAV>-<32540>-<212656>-<>-<>-<>-<
...

[Operators]
1=GSF;14 jun 2004;08 jul 2004
2=CLC;14 jun 2004;24 jun 2004
3=MPB;08 jul 2004;13 jul 2004
4=GA;22 jul 2004;21 sep 2004

[Validation specifications]
Table=1
LeadingSil=0|0
TrailingSil=0|0
Orthographic Events=0
Phonetic Events=0
Fast quality labeling=0
Source table=0
Channel=1

[End]

E.3 Hasierako HMM prototipoa

```
~o <VecSize> 39 <MFCC_0_D_A>
~h "proto"
<BEGINHMM>
 <NUMSTATES> 5
 <STATE> 2
   <MEAN> 39
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0
   <VARIANCE> 39
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0
 <STATE> 3
   <MEAN> 39
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0
   <VARIANCE> 39
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0
 <STATE> 4
   <MEAN> 39
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0
   <VARIANCE> 39
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0
 <TRANSP> 5
   0.0 1.0 0.0 0.0 0.0
   0.0 0.6 0.4 0.0 0.0
   0.0 0.0 0.6 0.4 0.0
   0.0 0.0 0.0 0.7 0.3
   0.0 0.0 0.0 0.0 0.0
<ENDHMM>
```