Euskarazko Speech_Dat (II) Datu-basea: Deskribapena eta Lehen Ezagutze Saiakeren Emaitzak

Inma Hernáez, Iker Luengo, Eva Navas, Maren Zubizarreta, Iñaki Gaminde, Jon Sanchez eta Imanol Madariaga

Elektronika eta Telekomunikazioak Saila
Euskal Herriko Unibertsitatea
inma, ikerl, eva, maren, igaminde, ion, imanol@bips.bi.ehu.es

Hitz gakoak: Hizkuntzaren teknologiak

Abstract

In this work we present a telephone speech database for Basque, compliant with the guidelines of the Speechdat project. The database contains 1060 calls from the fixed telephone network. We first describe the main aspects of the database design. We also present the recognition results using the database and a set of procedures following the language independent reference recogniser commonly named Refrec.

Laburpena

Lan honetan Speechdat proiektuaren jarraibideekin bat datorren euskararako telefono ahots datu-basea aurkezten dugu. Datu-baseak telefono finkoko saretik egindako 1060 dei ditu. Hasteko datu-basearen diseinuaren ezaugarri nagusiak azaltzen ditugu. Datu-basea eta Refrec deituriko hizkuntzaren menpe ez dagoen erreferentziazko ezagutzailearen prozedura multzoa jarraituz lortutako errekonozimenduaren emaitzak aurkezten ditugu ere. [1]

1. Sarrera

Hizkuntzaren teknologien garapenak, aplikazio, tresna eta ikerketen zimentarria diren hizkuntza baliabideak prestatu beharra dakar. IXA taldeak hizkuntza idatziaren tratamendua jorratu duen bitartean [1], hemen ahozko arlorako prestatu dugun baliabide bat aurkezten da.

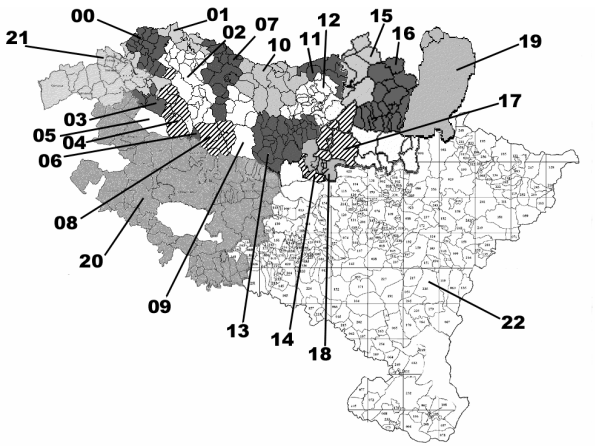
Artikulu honetan Euskararako telefono ahots datu-base baten prestaketa azaltzen dugu, euskaren ezagutzan emaitza batzuk lortzeko helburuarekin. Datu-baseak SpeechDat(II) jarraibideak betetzen ditu [2]. Euskalki eta eskualdeen inguruko erabakiak 2.atalean azaltzen dira. 3. atalean datu-basearen bilketa prozesua azaltzen da. 4. atalean egindako ezagutze esperimentuak azaltzen dira.

2. Hizlarien informazio demografikoa

2.1. Eskualde dialektalak

Aldaera dialektal asko dagoen arren, soinu multzoa ia berdina da kontuan hartu ditugun eskualde guztietan (hau da, iparraldeko euskalkiak alde batera utzita), Bizkaierako eskualde batzuetan agertzen den "Z" sabaiaurreko frikari ahostuna izan ezik [3]. Horregatik euskalkiarekiko eskualde bi bereizi ditugu. Batak, *Batua* deitu dugunak, euskal hiztunen %75a hartzen du barnean, baita Bizkaiera izan ezik beste euskalki guztiak eta batua ere. Bigarrenak, hiztunen beste %25a barneratzen duenak, "Z" soinua daukaten Bizkaiko eskualdeak ditu eta *Bizkaiera* deitu dugu. Lau hiri nagusiak lehen eskualdean daude.

Eskualde bakoitzean azentu ezberdinak hartu dira kontuan, aldaketa posible guztien bilketa ziurtatzeko. Batuan 13 azentu eredu definitu ziren eta Bizkaiera 10 azentutan sailkatu zen, guztira 23 azentu talde daude



1.Irudia: Euskara FDB1000 datu-basean zehaztutako 23 azentu eskualdeen irudikapena. 00-tik 09-rako eskualdeak Bizkaiera taldea osatzen dute, besteek Batua.

[1] Lan honek Espainiako MCYT-en dirulaguntza izan du TIC2000-1005-C03-03 proiektupean eta Euskal Herriko Unibertsitatearena UPV00147.345-E-14895/2002 proiektuan

datu-basean. Hauetariko azentu bakoitza eskualde geografiko batekin lotuta dago, 1. irudiko mapan agertzen den bezala.

Beraz kontu bereziagaz aukeratu da hizlarien jatorri geografikoa, datu-baseko azentu banaketa eta populazioaren arteko benetako banaketaren artean parekatze onena lortzeko asmotan. 1996-ko urtarrileko zentsuko datuak erabili dira horretarako [4] [5].

1300 formulario sortu ziren banatzeko. 300 Bizkaiera taldean dauden eskualdeetarako, euskalkiaren material berezia erabiliz, eta Batua talderako euskara batuaz prestatutako 1000 formulario erabili ziren. Askeneko datu-basean, 2060 sesiotatik, 273 Bizkaierako hizlarietatik grabatu ziren eta besteak, 787, Batuan.

Hizlarien azentua beraien jatorri geografikoaren arabera sailkatu zen. Ahotsaren azentuarekiko bigarren hezkuntzako aldia da erabakigarriena [6]. Beraz azentu sailkapena hizlariaren 14 eta 16 urte arteko bizilekuaren arabera egin zen. Informazio hau zuzenean grabaketetatik lortu zen, hizlariei deian bertan galdetzen zitzaien eta.

2.2. Adin eta genero banaketa

1. taulan Euskara FDB1000 datu-baseko hizlarien banaketa agertzen da, adin eta generoaren arabera. SpeechDat(II)-ko espezifikazioak [6] betetzen dituela ere ikusten da.

	Emak.	**Gizon.**	**Guzt.**	**%**	***Espezif.***
<16	5	3	8	**0,75**	*1% gomen.*
16-30	265	209	474	**44,72**	*>20%*
31-45	185	135	320	**30,19**	*>20%*
46-60	116	120	236	**22,26**	*>15%*
>60	6	7	13	**1,23**	*Aukeran*
Ezezag.	3	6	9	**0,85**	
Guztira	580	480	1060	**100**	
%	**54,72**	**45,28**	**100**		
Espezif.	*%45-55*	*%45-55*			

1.Taula: Hizlarien banaketa adin eta generoaren arabera

3. Datu-basearen bilketa

3.1. Grabaketarako tresneria

PCan oinarritutako grabaketa sistema garatu zen, RDSI sarera lotutako interfaze txartelarekin. Seinale fitxategiak Windows 98, AVM-RDSI txartela eta UPCan garatutako ADA softwarea [7] erabiliz grabatu ziren. Fitxategiak disko gogorrean zuzenean gordetzen ziren eta aldizka segurtasun kopia egin. PC bat erabili zen, gehienez dei bi batera erantzun ahal zituelarik.

Fonemak irudikatzeko SAMPA kodea erabili zen, aipatutako Bizkaierako "Z" barneratzen duen 2. taulan agertzen den Euskararako SAMPA [8].

3.2. Hizlariak lortzeko

Hizlariak lortzeko 20 laguntzaile boluntarioren bitartez egin zen. Hauek lortutako dei bakoitzeko diru pixka bat jasotzen zuten, 1'2 eta 3'6 € bitartean, dei kopuruaren arabera — dei gehiagorekin diru gehiago.

Partehartzea bultzatzeko hizlariek 200 € irabazi zezaketen zozketa egin zen.

Deitu behar zuten hizlariek laguntzaileen bidez jasotako azalpen eta galdetegi formularioa sinatzen zuten. Laguntzaileek formularioak jaso eta bueltatu ondoren kobratzen zuten jasotako deiengatik, eta hizlariek zozketan sartzen ziren.

Hizlari gehiago lortzeko pertsonaz pertsona egindako harremanak ere erabili ziren formulario eta galdetegi gehiago banatzeko.

Ikur	Berba	Trans.	Azalpena
Leherkari			
p	apeza	apes`a	ezpainetako leherkari ahoskabea
b	begia	beGia	ezpainetako leherkari ahostuna
t	etorri	etorri	hortzetako leherkari ahoskabea
c	ttantta	canca	sabaikari leherkari ahoskabea
d	denda	denda	hortzetako leherkari ahostuna
k	ekarri	ekarri	leherkari belare ahoskabea
g	gaia	gaia	leherkari belare ahostuna
Afrikatuak			
tS	txikia	tSikia	afrikatu sabaikari ahoskabea
ts	atso	atso	afrikatu bizkar-hobietako albeolar ahoskabea
ts`	atzo	ats`o	goi-hobietako afrikatu ahoskabea
gj	onddo	ongjo	afrikatu sabaikari ahostuna
Frikariak			
jj	leoia	leojja	frikari sabaikari ahostuna
f	afaria	afaria	espain-hortzetako frikari ahoskabea
B	hamabi	amaBi	ezpainetako hurbilkari ahostuna
T	perez	pereT	hortzarteko frikari ahoskabea
D	adarra	aDarra	hortzetako hurbilkaria ahostun
s	hasi	asi	bizkar-hobietako frikaria ahoskabea
s`	zoroa	s`oroa	goi-hobietako frikari ahoskabea
S	xoxoa	SoSoa	sabaiaurreko frikari ahoskabea
x	ijito	ixito	frikari belare ahoskabea
G	agur	aGurr	frikari belarea ahostun hurbilkaria
Z	berria	berriZa	sabaiaurreko frikari ahostuna
Sudurkari			
m	ama	ama	ezpainetako sudurkari ahostuna
n	neska	neska	hobietako sudurkari ahostuna
J	ñabar	JaBarr	sabaiko sudurkari ahostuna
Liquids			
l	lana	lana	hobietako saiheskari ahostuna
L	iluna	iLuna	sabaiko saiheskari ahostuna
r	dirua	dirua	hobietako dardarkari bakun ahostuna
rr	arrunta	arrunta	hobietako dardarkari anitz ahostuna
Bokalak			
i	ipar	iparr	aurreko bokal itxi ezbiribila
e	hemen	emen	erdiko bokal itxi ezbiribila
a	ama	ama	erdiko bokal ireki ezbiribila
o	oso	oso	atzeko bokal erdi borobila
u	umore	umore	atzeko bokal itxi borobila

2.Taula: Euskara SpeechDat datu-basean erabilitako SAMPA ikurrak, adibideekin.

3.3. Hizlari motak sailkatzen

Hizlariei buruz informazio hau gordetzen zen:

- Grabaketaren eguna eta ordua automatikoki.

- Generoa, adina, izena eta helbidea jasotako galdetegitik.

- Grabaketaren baldintzak, euskalkia eta ingurumena ahozko galderetatik.

Deiaren eskualdea jakiteko, automatikoki gordetako telefono zenbakia erabiltzen zen “alderantzizko telefono gida” programa baten bidez posta kodea lortzeko.

.Test	Ezagutze lana
I	Zenbaki isolatuak
Q	Bai/Ez
A	Aplikazio berezietarako hitzak
BC	Zenbaki lotuen zerrenda
O	Hiri izenak
W	Berba fonetikoki aberatsak

3.Taula: Test arruntak eta beraien azpi-korpusak

4. Errekonozimenduaren emaitzak

COST 249 proiektu Europarraren [9] helburuetariko bat ahots ezagutzaileen entrenamendu eta saiakerarako erreferentzi prozedurak ezartzea zen, hizkuntzarekiko dependentzia txikienarekin. Ondorioz hizkuntzaren menpe ez dagoen erreferentziazko ezagutzailea garatu zen, *Refrec* deiturikoa. Ezagutzaile honek fonema ereduen multzoa entrenatzen du SpeechDat(II)-ren baldintzak betetzen dituen edozein datu-basetik zuzenean, datu-basean dagoen hizkuntzaren menpeko informazioa erabiliz. Aurre-segmentatutako datu barik lan egiten duen prozedura erabiltzen du, HTK lanabesarekin [10]. Era honetan hizkuntza ezberdinetarako konparatu daitezken ezagutze emaitzak lor daitezke.

Erreferentziazko ezagutzailea entrenatzeko prozedura [10] liburuan HTK ikasteko adibidetik garatu da. Hiru estatuko eskerretik-eskumarako Markoven eredu ezkutua (HMM) transkripzio ortografikoetatik eta ebakera SpeechDat(II) datu-basean emandako lexikoitik entrenatu da.

Ezaugarri akustikoak 39 dimentsiotako Mel cepstral koefizienteak dira (MFCC), zero koefiziente cepstrala energia bezala, lehen eta bigarren deltak barne.

Ereduek hizkuntza ezberdinetan daukaten portaera aztertzeko, SpeechDat(II) datu-basean bakarrik oinarritutako test batzuk diseinatu ziren. Helburu horretarako SpeechDat(II)-ren sesio ofizialak erabili ziren. Sei test diseinatu ziren azpi-korpus batzuetarako, 3. taulan agertzen denez.

Euskara Speech_Dat(II) datu-basea erabili zen ezagutzaile hau entrenatzeko, eta sei testak pasarazi. Errekonozimenduaren emaitzak Refrec-en web ofizialean [11] publikatuta dauden beste hizkuntzatan lortutakoekin konparatu dira.

4. taulan agertzen dira hizkuntza hauen estatistikak, Euskara barne. Balio hau konparatuz argi dago Euskara dela fonema kopuru txikiena daukan hizkuntza. Beste hizkuntza batzuetan (Daniera adibidez) diptongoak fonema bezala erabiltzen dira lexikoian, entrenatutako ereduen kopurua handituz. Baina entrenatutako trifonema ereduen kopurua handiagoa da Euskara Speech_Dat(II) datu-basean, datu-base handiagoek bakarrik (2000, 4000 eta 5000 sesiokoek) gainditzen dutena. Beraz Euskara SpeechDat datu-basean

grabatutako esaldiek fonema trantsizio ezberdinen kopuru handiagoa daukate, eta horregatik ezagutze sistemak diseinatzeko informazio fonetiko gehiago daukate, datu-basearen neurri berdinerako. Datu-basearen diseinuan, trantsizio fonetikoen kopuru handiena lortzea bilatu zen, batez ere S eta W azpi-korpusen edukinean. Bestalde ikusten da Euskarazko trifonemak ez dakartela pilaketaren bidezko estatu gutxitxe handia. pilaketari begira, ikus daiteke Euskarako trifonemak ez dituztela estatuak asko gutxitzen.

Errekonozimenduaren emaitzak, test bakoitzak lortutako berba errore tasa txikiena (WER), 5. taulan agertzen dira. Taula hau 6. taularekin batera aztertu behar da, berbategiak daukan berba kopurua eta batezbesteko fonema kopurua erakusten duena. Espero zen bezala, emaitzak berbategi txikiko testetan erdi-neurriko berbategikoetan baino hobeak dira. Baina ezberdintazun nabarmenak daude hizkuntza ezberdinen artean. Erdi-neurriko testetan (O eta W testak), berbategiaren neurriren ezberdintasunek ondorio nabarmena daukatela dirudi. Berbategi txikiko testen ezberdintasunak azaltzea zailagoa da. Batzuk telefono sareetako zaratarekin lotuta egon daitezke [12]. Berbategi eta fonema multzoetako ezberdintasunak ere eragina izan dezake.

5. taulan agertzen dan moduan, Euskara FDB1000 datu-basea kasu bakoitzeko onenen artean dago, datu-base eta berbategiaren neurria kontuan hartu ezkero.

5. Aipamenak

[1] Alegria, I; Artola, X.; Sarasola, K.; “Hizkuntzaren tratamendu automatikoa: aplikazioak, tresnak, baliabideak eta oinarriak” Euskalingua 2002, 1, pp 86-90.

[2] LE2-4001 SpeechDat (II) proiektuaren web orria http://www.speechdat.org/SpeechDat.html

[3] Gaminde, I. “Bizkaiko euskararen ezaugarri fonologiko batzuen inguruan” Euskalingua 2002, 1, pp 4-14.

[4] EUSTAT-Basque Statistical office, Main results of Population and Housing Statistics 1996. Language Census, www.eustat.es/english/general/pob_viv/pob_vi.html

[5] Navarre Statistical Institute, www.cfnavarra.es/estadistica/

[6] SpeechDat Deliverable LE-4001-SDI1.2.1 version 2.2, “Enviromental and speaker specific coverage for Fixed Networks”, Feb. 1997

[7] José A. R. Fonollosa, A. Moreno, “Automatic Database Acquisition Software for ISDN PC Cards and Analogue Boards”, Proc. of LREC, Granada (Spain), May 1998, pp 28-30.

[8] Zalbide, X.; Gaminde, I.; Hernaez, I.; Zubizarreta, M.; Navas, E., 2003 "Euskararako SAMPA kodeaz" Euskalingua 2003, 2, pp.171-177.

[9] F.T. Johansen, N. Warakagoda, B. Lindberg, G. Lehtinen, Z. Kacic, A. Zgank, K. Elenius and G. Salvi, "The COST 249 SpeechDat Multilingual Reference Recogniser", Proc. of LREC, Athens, May 2000, Vol 3, pp 1351-1355.

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, “The HTK Book (for HTK Version 3.0)”, Cambridge University, Cambridge, England, 2000.

[11] COST 249 SpeechDat SIG, 2000 "The Refrec homepage", www.telenor.no/fou/prosjekter/taletek/refrec

[12] Børge Lindberg, “The Danish SpeechDat(II) Corpus - A Spoken Language Resource”, DALF Proceedings, Centre for Language Technology, Copenhagen, May, 1999.

Hizkuntza (datu-basea)	Entr n. ses.	Lexikoi ebak.	Mono-fonem	Esald. gehi.	Entren esal.	Tri-fon	Estatu pilak. gutx.
Daniera FDB1000	800	39604	71	34400	23216	13056	7.3 %
Daniera FDB4000	3500	39604	71	150500	101100	19032	11.5 %
Nederlandera	4522	–	47	22602	20167	10194	8 %
Ingelesa FDB1000	866	12149	44	39831	27374	8060	11.2 %
Ingelesa MDB1000	800	–	43	30917	26068	8368	–
Aleman FDB1000	860	23578	47	37335	24158	11472	9.3 %
Norvegiera FDB1000	816	14826	40	36720	20335	7866	8.4 %
Esloveniera FDB1000	800	6011	39	34392	20548	6613	10.8 %
Suediera FDB1000	800	25946	46	38400	24827	10689	8.6 %
Suediera MDB1000	800	16050	46	41600	34346	11876	7.8 %
Suediera FDB5000	4463	65675	46	214223	179807	16009	15.9 %
Suitzako Alemana FDB1000	800	30525	51	32580	17442	12374	7.1
Suitzako Alemana FDB2000	1500	49713	45	61055	37675	14229	9.5 %
Euskara FDB1000	860	48273	33	36980	28677	14091	18.2%

4. Taula: Entrenamendu estatistikak.

Hizkuntza (datu-basea)	Test korpusa					
	I	**Q**	**A**	**BC**	**O**	**W**
Daniera FDB1000	1.0	1.1	2.4	2.3	15.8	64.4
Daniera FDB4000	0.6	1.1	2.4	2.7	14.0	64.1
Alemanera	–	–	–	5.0	–	–
Ingelesa FDB1000	2.6	0.4	1.4	4.3	6.0	34.3
Ingelesa MDB1000	10.2	–	–	–	–	–
German FDB1000	0.8	0.0	2.4	2.7	6.0	8.7
Norvegiera FDB1000	2.3	0.5	4.4	5.9	17.3	34.7
Esloveniera FDB1000	4.2	0.9	4.9	6.1	9.3	19.3
Suediera FDB1000	1.0	0.0	1.2	2.5	12.4	35.2
Suediera MDB1000	10.5	1.1	4.0	14.2	18.6	52.4
Suediera FDB5000	2.6	0.7	2.5	4.5	21.3	79.9
Suitzako Alemana FDB1000	0.5	0.3	1.1	3.1	6.3	24.3
Suitzako Alemana FDB2000	0.0	0.8	0.6	2.4	9.6	33.4
Euskara FDB1000	0.0	0.0	1.0	1.8	8.2	17.2

5. Taula: Refrec0.95-en lortutako berba errore tasak (%).

Hizkuntza (datubasea)	I/BC		Q		A		O		W	
	Berba#	**Fnmk**	**Berba#**	**Fnmk**	**Berba#**	**Fnmk**	**Berba#**	**Fnmk**	**Berba#**	**Fnmk**
Daniera FDB1000 Daniera FDB4000	11	2.64	2	2.00	30	4.57	495	6.52	16934	8.76
Ingelesa FDB1000	10	2.87	2	2.50	31	4.90	259	8.04	2527	5.50
Alemanera FDB1000	10	3.40	2	2.50	30	6.30	374	7.67	2264	10.71
Norvegiera FDB1000	10	2.85	2	2.00	30	4.60	1182	7.34	3438	6.59
Esloveniera FDB1000	10	3.85	2	2.00	31	6.52	597	10.36	1491	6.75
Suediera FDB1000	10	3.33	2	2.50	30	6.23	905	9.29	3610	9.31
Suediera MDB1000	10	3.33	2	2.50	30	6.23	869	8.96	3611	9.13
Suediera FDB5000	10	3.33	2	2.50	30	6.23	2344	11.07	18249	8.75
Suitzako Alemana FDB1000	10	3.70	2	2.50	30	6.67	684	12.64	3274	7.90
Suitzako Alemana FDB2000	10	3.70	2	2.50	30	6.67	1218	12.97	5319	7.87
Euskara FDB1000	10	3.90	2	2.25	30	6.40	768	8.32	3968	8.22

6.Taula: Berba kopurua eta batezbesteko fonema kopurua berba bakoitzeko test berbategietan.