The Basque Speech_Dat (II) Database:
A Description and First Test Recognition Results

I. Hernaez, I. Luengo, E. Navas, M. Zubizarreta, I Gaminde, J. Sanchez

Department of Electronics and Telecommunications
University of th Basque Country
`inma, ikerl, eva, maren, igaminde, ion@bips.bi.ehu.es`

Abstract[1]

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.

1. Introduction

Basque language is official in the Autonomous Basque Community (Spanish area of the Basque speaking region). It is possible to study in Basque in all levels of education, although graduate level is limited to a small number of curricula, and post-graduate studies are almost non-existent. According to data extracted from [1] and [2] there are almost one million speakers of Basque, including those that may use mainly Spanish in every day use, but are able to communicate in Basque and do it occasionally.

The language has a very high dialectal fragmentation with 7 main dialects (4 of them present in the Spanish area) and more than 50 varieties according to modern commonly accepted assumptions. A standardization process of written Basque started on 1975, unifying and regulating the lexicon, the grammar, the syntax, the morphology and every other aspect of the written language. From many points of view, this standard Basque adopted many of the features from one of the most spoken dialects, the so called *Guipuscoan* dialect. As for pronunciation, no official rules have been given.

In this work we describe the creation of a telephone speech database for Basque, with the purpose of obtaining some recognition results for this language. The database follows the SpeechDat (II) guidelines [3]. The decisions taken concerning dialects and regions will be explained in Section 2, together with other speaker information in the database. Section 3 describes the database collection process. Section 4 is devoted to the description of the recognition experiments performed.

2. Speaker demographic information

2.1. Dialectal regions

In spite of the huge number of dialectal varieties, the set of sounds is almost the same for the considered regions (i.e. leaving aparte the northern dialects), with the exception of the voiced pre-palatal fricative "dZ", which may occur in some areas of the Biscayan dialect speaking area. This is why we distinguished two main dialectal regions. The first one would cover 75% of the Basque speaking population, including all the dialects except for the Biscayan dialect, as well as the standard variety, which has given the name to the dialectal region, *Batua* (which is the basque word for the standard Basque) The second one, covering the remaining 25%, would include the Biscayan dialects where the dZ sound was expected, and it was name as the dialect, *Bizkaiera*. All the four main cities of the area where included in the first region.

In each of the two regions, different accents were considered, to assure the collection of every possible variant in the database. 13 accents types were defined for Batua whereas Bizkaiera was classified in 10 accents, giving a total of 23 different accent groups in the database. Each of these accents is associated to a certain geographical region, as shown on the map of Figure 1.

As such, special care was taken when selecting the speakers geographical origin, in order to achieve the best agreement between the accent distribution within the database and the actual distribution of the population. Census data of January 1996 was used for this purpose [1], [2].

A total of 1300 prompt sheets were created for distribution. 300 of them were assigned to regions belonging to the Bizkaiera group, using specific material of the dialect, while standard Basque was used for the Batua group (1000 prompt sheets). In the final database, from a total of 1060 sessions, 273 were recorded from Bizkaiera speakers, while the remaining 787 were carried out in Batua.

The speakers accent was classified according to their geographical origin. As stated in [4], the high school period is most decisive for the accent in the speech. Therefore, the accent classification was carried out in accordance with the speaker's region of residence at the age of 14 or 16. This information was directly obtained from the recordings, since speakers were asked during the call.

2.2. Age and gender distribution

Table 1 shows the distribution of speakers in the database, in terms of gender and age. The SpeechDat(II) specifications defined in [4] are also presented, showing that the Basque FDB1000 database agrees with the requirements.

3. Database collection

3.1. Recording platform

In order to fulfill the requirements, a PC-based recording system with an ISDN interface board connected to a local exchange was developed. Signal files were recorded with a
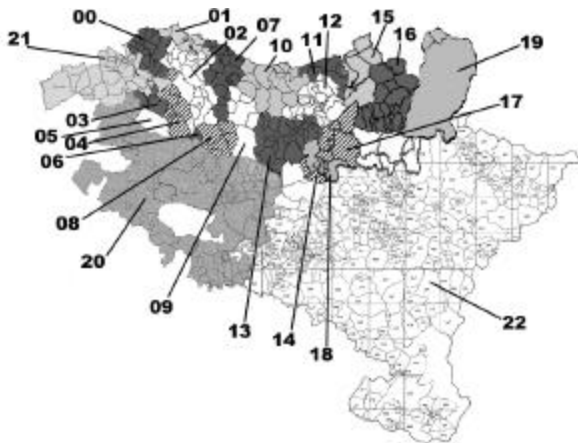

Figure 1: Representation of the 23 accent regions defined within the Basque FDB1000 database. Regions from 00 to 09 form the Bizkaiera group. The rest represent Batua accents.

	Female	**Male**	**Tot.**	**%**	*Specif.*
<16	5	3	8	**0,75**	*1% recom.*
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**	*Optional*
Unkn	3	6	9	**0,85**	
Total	580	480	1060	**100**	
%	**54,72**	**45,28**	**100**		
Specif.	*45-55%*	*45-55%*			

Windows 98 based system using an AVM-ISDN board and the ADA software developed by UPC [5]. Files were stored directly onto hard disk and backed up regularly. One PC was used giving a maximum capacity of 2 parallel calls.

3.2. Speaker recruitment

Speakers were recruited indirectly, by means of 20 voluntary recruiters who were given a small cash incentive for each call they obtained. The incentive was between 1,2 and 3,6 € depending on the number of calls — the more calls they got, the more money they earned.

Participation was encouraged with a draw in which speakers could win 200 €

Prompt sheets with instructions and questionnaires to be signed by callers were given to recruiters who distributed them to speakers. Recruiters took care of re-collecting the questionnaires, and only when they returned them properly signed were paid for the received calls, and the corresponding callers entered in the prize draw.

As a further attempt to recruit speakers, person-to-person contact was used to distribute additional prompt sheets and questionnaires.

3.3. Gathering speaker typology

The following information has been collected from the speakers:
- Date and time of recording are automatically logged.
- Gender, age, name and address from the returned questionnaire.
- Recording conditions, dialect and environment are gathered from the spontaneous questions.

For the region of call, the automatically logged telephone number was used to obtain the postal code through a "reverse phone book" program.

4. Recognition results

One of the purposes of the COST 249 European project [6] was the definition of reference procedures for training and testing speech recognisers, with a minimal dependency on the language. As a result a language independent reference recogniser was developed, commonly named Refrec. This recogniser trains a set of phoneme models directly from any SpeechDat(II) compliant database using the language-dependent knowledge embedded in the database, and relays on a boot-strapping procedure that works without pre-segmented data, based on the HTK toolkit [7]. In this way it is possible to get comparable recognition results for different languages.

The reference recogniser training procedure is an extension of the HTK tutorial example in [7]. Three state left-to-right hidden Markov model (HMM) monophone and word-internal state-clustered triphones (context sensitive phone models) with 32 Gaussian mixture components are trained from orthographic transcriptions and the pronunciation lexicon provided in the SpeechDat(II) database.

The acoustic features are 39-dimensional Mel cepstral coefficients (MFCC), including the zero'th cepstral coefficient as energy, first and second deltas.

In order to analyse the general behaviour of the models in different languages, some tests were designed based only on the SpeechDat(II) database itself. For this purpose, the official SpeechDat(II) test sessions are used. Six common test have so far been designed for some of the sub-corpora, as shown in Table 2.

The Basque Speech_Dat(II) database was used to train this recogniser, and each of these tests were applied. Recognition results have been compared with those obtained for other languages, which are published in the Refrec official web page in [8].

Table 4 shows the training statistics for these languages, including Basque. Comparing these values, it becomes clear that Basque is the language with the minimum number of phonemes. In other languages (such as in Danish) diphthongs are treated as phonemes in the lexicon, raising the number of trained models. Nevertheless, the number of trained triphone models is greater in the Basque SpeechDat(II) database, only exceeded by larger databases (with 2000, 4000 and 5000 sessions). This means that utterances recorded in the Basque SpeechDat database present a larger number of different phone transitions, and thereby, present more phonetic information for the design of recognition systems, maintaining the size of the database. During the database's design, care was taken to maximize the number of phonetic transitions, specially in the contents of the S and W sub-corpora. On the other hand, analysing the state cluster reduction performance, it can be seen that Basque triphones yield to a limited state reduction.

Table 2: Common tests an associated sub-corpora.

Test	Recognition task
I	Isolated digits
Q	Yes/No
A	Application words
BC	Connected digit strings
O	City names
W	Phonetically rich words

Recognition results are shown in Table 5, in terms of the minimum word error rate (WER) achieved for each test. This table should be analysed simultaneously with Table 6, which presents the number of words and the average number of phonemes per word in the vocabulary. As expected, results in small vocabulary tests are clearly better than in medium vocabulary tests. But there is a considerable difference between languages. For medium vocabulary tests (O-test and W-test), the difference in the size of the vocabulary seems to have a significant impact on the results. Differences in the small vocabulary tests are harder to explain, although some of these can be related to the different noise level in the telephone networks [6]. Differences in the vocabulary and phoneme sets contribute significantly as well.

As shown in Table 5, test results for the Basque FDB1000 database are among the best ones on each case, if the size of the database and the vocabulary are taken into account.

5. References

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

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

[3] LE2-4001 SpeechDat (II) project homepage
http://www.speechdat.org/SpeechDat.html

[4] SpeechDat Deliverable LE-4001-SDI1.2.1 version 2.2, "*Enviromental and speaker specific coverage for Fixed Networks*", Feb. 1997

[5] 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.

[6] 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.

[7] 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.

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

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

Table 3: SAMPA symbols used in the Basque SpeechDat database, with examples.

Simb.	Word	Trans.	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	adarra	aDarra	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	agur	aGurr	voiced velar fricative approximant
Z	berria	berriZa	voiced prepalatal ficative
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
Vowels			
i	ipar	iparr	vowel front close unrounded
e	hemen	emen	vowel front mid unrounded
a	ama	ama	vowel central open unrounded
o	oso	oso	vowel back mide rounded
u	umore	umore	vowel back close rounded

Table 4: Training statistics.

Language (database)	Train ses.	Lexicon pronuns	Mono-phns	Max uttr.	Train uttr.	Tri-phones	State clstr red.
Danish FDB1000	800	39604	71	34400	23216	13056	7.3 %
Danish FDB4000	3500	39604	71	150500	101100	19032	11.5 %
Dutch	4522	–	47	22602	20167	10194	8 %
English FDB1000	866	12149	44	39831	27374	8060	11.2 %
English MDB1000	800	–	43	30917	26068	8368	–
German FDB1000	860	23578	47	37335	24158	11472	9.3 %
Norwegian FDB1000	816	14826	40	36720	20335	7866	8.4 %
Slovenian FDB1000	800	6011	39	34392	20548	6613	10.8 %
Swedish FDB1000	800	25946	46	38400	24827	10689	8.6 %
Swedish MDB1000	800	16050	46	41600	34346	11876	7.8 %
Swedish FDB5000	4463	65675	46	214223	179807	16009	15.9 %
Swiss German FDB1000	800	30525	51	32580	17442	12374	7.1
Swiss German FDB2000	1500	49713	45	61055	37675	14229	9.5 %
Basque FDB1000	860	48273	33	36980	28677	14091	18.2%

Table 5: Word error rates (in %) achieved on Refrec0.95.

Language (database)	Test corpus					
	I	Q	A	BC	O	W
Danish FDB1000	1.0	1.1	2.4	2.3	15.8	64.4
Danish FDB4000	0.6	1.1	2.4	2.7	14.0	64.1
Dutch	–	–	–	5.0	–	–
English FDB1000	2.6	0.4	1.4	4.3	6.0	34.3
English MDB1000	10.2	–	–	–	–	–
German FDB1000	0.8	0.0	2.4	2.7	6.0	8.7
Norwegian FDB1000	2.3	0.5	4.4	5.9	17.3	34.7
Slovenian FDB1000	4.2	0.9	4.9	6.1	9.3	19.3
Swedish FDB1000	1.0	0.0	1.2	2.5	12.4	35.2
Swedish MDB1000	10.5	1.1	4.0	14.2	18.6	52.4
Swedish FDB5000	2.6	0.7	2.5	4.5	21.3	79.9
Swiss German FDB1000	0.5	0.3	1.1	3.1	6.3	24.3
Swiss German FDB2000	0.0	0.8	0.6	2.4	9.6	33.4
Basque FDB1000	0.0	0.0	1.0	1.8	8.2	17.2

Table 6: Number of words and average number of phonemes per word in the test vocabularies.

Language (database)	I/BC		Q		A		O		W	
	#Word	Phns	#Word	Phns	#Word	Phns	#Word	Phns	#Word	Phns
Danish FDB1000 **Danish FDB4000**	11	2.64	2	2.00	30	4.57	495	6.52	16934	8.76
English FDB1000	10	2.87	2	2.50	31	4.90	259	8.04	2527	5.50
German FDB1000	10	3.40	2	2.50	30	6.30	374	7.67	2264	10.71
Norwegian FDB1000	10	2.85	2	2.00	30	4.60	1182	7.34	3438	6.59
Slovenian FDB1000	10	3.85	2	2.00	31	6.52	597	10.36	1491	6.75
Swedish FDB1000	10	3.33	2	2.50	30	6.23	905	9.29	3610	9.31
Swedish MDB1000	10	3.33	2	2.50	30	6.23	869	8.96	3611	9.13
Swedish FDB5000	10	3.33	2	2.50	30	6.23	2344	11.07	18249	8.75
Swiss German FDB1000	10	3.70	2	2.50	30	6.67	684	12.64	3274	7.90
Swiss German FDB2000	10	3.70	2	2.50	30	6.67	1218	12.97	5319	7.87
Basque FDB1000	10	3.90	2	2.25	30	6.40	768	8.32	3968	8.22