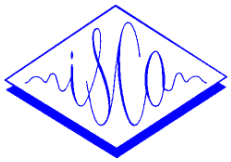

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SELECTION OF SUBLEXICAL UNITS FOR CONTINUOUS SPEECH RECOGNITION OF BASQUE.

M.K. López de Ipiña[1], I.Torres[2],L.Oñederra[3], A. Varona[2], L.J. Rodríguez[2].

[1]Sistemen Ingeniaritza eta Automatika Saila. Gasteiz.
[2] Elektrika eta Elektronika Saila. Bilbo.
[3] Euskal Filologi Saila. Gasteiz .

University of the Basque Country. Spain.

email: karmele@we.lc.ehu.es

ABSTRACT

This paper describes the work carried out to select the most suitable set of Sublexical Units for Continuous Speech Recognition of Basque. Even if there are several dialects in Basque, only one of them has been used to choose the preliminary set of sounds. Bearing in mind this aim, a wide experimentation has been carried out to select Context Independent Phone-Like Units. Then, in order to obtain robust acoustic models for the language, the units have been evaluated with most of the dialectal variants of Basque. Finally, Decision-Trees based Context Dependent Sublexical Units are selected. For building the trees the classical methodology of Bahl and the efficient Growing and Pruning algorithm have been used.

1. INTRODUCTION

The selection of appropriated sublexical units is very important for a Continuous Speech Recognition (CSR) system. The first step in this selection is the choice of the basic subset of Context Independent Phone-Like Units (CI-PLUs). These units are chosen among the wide variety of sounds of the language.
Basque has about 1.000.000 speakers in Basque Country and presents a wide dialectal distribution being eight the main dialectal variants. This dialectal variety involves differences at phonetic and phonologic levels being different the fundamental set of sounds for some of the variants. Therefore, a previous selection of the widest set of sounds among the dialects has to be carried out in order to choose the set of CI-PLUs. Afterwards the modelling of context dependencies could improve the performance of the these units. Thus, the classical methodology of Decision Tree clustering will be used to chose Decision-Tree based Context Dependent Units (DT-CDUs)
Next section describes the allophonic variety of the Basque language. In this work, the used database is presented in section 3. Section 4 is devoted to the selection of CI-PLUs. Section 5 describes the selection of DT-CDUs. Finally, conclusions are remarked in Section 6.

This work has been partially supported by Spanish CICYT under grant TIC98-0423-C06-03 and GV/PI98/111

2. PHONETIC ANALYSIS OF THE LANGUAGE.

Althought the Basque language is spoken by a reduced community, it presents a wide dialectal distribution due to its historical geographical isolation and old sociological and historical-political factors [1]. In fact there are eight main dialectal variants. This diversity involves differences at phonetic, phonologic and morphological levels. Specially, there are differences among dialects in the set of sounds in the both phonetic and phonological levels.

Regarding the dialects of the south of the Basque Country the differences emerge in sibilants as in the utterance of the africates [ts] [ts_a] and the fricatives [s] [s_a], where [s_a] and [ts_a] are apical. These sounds differ from some dialects (Guipuzcoan and Navarrese) to others (Bizcayan). In the case of northern dialects there is an influence of French phonetics, specially in the pronunciation of the vowel [u] and the aspiration of the [h]. These dialects will be considered in future works.

It is also relevant the existence of the unified Basque, *Batua* an standardisation of the language which was officially instituted by the Basque Academy in 1968, it was created with the aim of overcoming the dialectal differences. The *Batua*, which is closer to some dialects than to others, has nowadays a big importance in the Basque community, being used both for a big number of persons whose mother language is not the Basque Language and in the public institutions, most of mass-media and in the text of generic subjects.

Thus, analysing the statistics of the phonetics of Basque several features can be observed (table 1):

1) The global maxima are situated in vowel sounds being dominants [e] and [a].
2) There are also some local maxia although of litter magnitude in the plosives, where dominate [t] and [k].
3) In nasals it can be seen a local maximum in [n].
4) With regard to the africates the maximum is situated clearly in [s] and the minima (global minima) are situated in sounds as [f] and [x].
5) Finally the africates have a local maximum in [ts] and a global minima in [ts_a] and [tS].

Note that with only five sounds it is achieved an Accumulated Frequency (AF) of the 50% (table 1). With half of the sounds the 90% of the text can be transcribed. With regard to the biphones, around 700 different ones were found in the analysis. These represent the 60% of the possible maximum (34 x 34 =1156).

3. THE DATABASE

In order to select an adequate set of Sublexical Units (SUs) for the southern variants, a new database was built, called EHBB[2]. In the design of the database just one of the dialects was used as reference for training: the *guipuzcoan*. This dialect presents the biggest set of sounds and it is the closest to *Batua*

Table 1. Inventory of *Guipuzcoan* sounds. Relative Frequency of Occurrence (RFO) and Accumulated Frequency (AF) of the sounds

Sound	RFO	AF
a-h**a**ndi**a**	16.77	16.77
e-**e**n**e**ko	12.79	29.56
i-b**i**de	7.54	37.1
n-**n**oa	6.55	43.65
o-h**o**ri	5.5	49.15
t-du**t**	5.49	54.64
u-**u**me	4.8	59.44
s-**zu**	4.27	63.71
k-to**k**i	4.27	67.98
rr-ha**rr**i	3.39	71.37
r-hu**r**a	2.66	74.03
j-be**hi**a	2.55	76.58
s_a-**s**u	2.17	78.75
m-a**m**a	1.89	80.64
S-mu**x**u	1.84	82.48
l-al**d**e	1.8	84.28
D-han**d**i	1.7	85.98
G-da**g**o	1.64	87.62
N-ha**n**ka	1.53	89.15
B-ho**b**e	1.4	90.55
b-en**b**at	1.3	91.65
c-an**tt**on	1.06	92.91
Ts-a**tz**o	1.02	93.93
w-ga**u**a	0.94	94.87
g-ain**g**eru	0.93	95.8
p-**p**attal	0.89	96.69
J-mi**n**a	0.74	97.43
L-mi**l**a	0.68	98.11
x-a**j**e	0.58	98.69
jj-**j**oan	0.48	99.17
tS-**tx**ori	0.42	99.59
f-**f**ede	0.17	99.76
ts_a-a**ts**o	0.12	99.88
D-ho**d**ei	0.12	100

.

The basic corpus for the spoken database consisted of 300 sentences that were selected according to the following criteria:

· The phonetic balance was guaranteed, thus the Relative Frequency of Occurrence (RFO) of each allophone in the corpus is the same as in the reference sample (table 1).

· A minimum number of occurrences of each allophone was considered in order to have enough samples for further stochastic modelling.

4. SELECTION OF CI-PLUSs

Based on the EHBB database a wide evaluation of different sets of CI-PLUs was carried out. Thus, the database was divided in three subsets: training, tuning and test. The training sub-set (TRAIN) consisted of 250 sentences uttered twice by 26 speakers. The Tuning set TUN consist of 50 sentences uttered twice by the same 26 speakers. The test sub-set was divided likewise in four subsets:

1) A Speaker-Dependent Vocabulary-Independent test (TSDVI), formed by 50 sentences uttered twice by 26 guipuzcoan speakers.

2) A Speaker Independent Vocabulary Dependent test (TSIVD), formed by 250 sentences uttered twice by 14 guipuzcoan speakers.

3) A Speaker Independent Vocabulary Independent test (TSIVI), formed by 50 sentences uttered twice by 14 guipuzcoan speakers

4) A Speaker Independent Vocabulary Independent test (TSIVING), formed by 50 sentences uttered twice by 20 no guipuzcoan speakers

The whole database consist of about 13,000 sentences (520,000 sounds) for training, 2,600 (104,000 sounds) for tuning, and 2,400 (96,000 sounds) for test.

Table 2. Statistics of the EHBB database. Number of Speakers (SPK), Sentences in the corpus , Total Number of Sentences (TST) and Total Number of Sounds (TSD)

	TRAIN	TUN	TSDVI	TSIVD	TSIVI	TSIVING
SPK	26	26	26	14	14	20
STN	250 *2	50*2	50*2	250*2	50*2	50
TST	13,000	2,600	2,600	7,000	1,400	1,000
TSD	520,000	104,000	104,000	280,000	56,000	40,000

Table 3. Recognition Rates of Preliminary experiment with CI-PLUs.

	N-UNITS	%REC	a	e	i	j	o	u	w	b	B	d	D	g	G	p	t	k	y	L	l	r	R	m	n	N	J	S	A	s	f	x	ts	tt	tS	ts
EXP1	34	61.85	74	76	75	79	72	78	64	61	57	58	58	44	42	72	70	63	57	84	59	66	73	75	48	76	82	93	80	83	86	93	52	72	50	75
EXP2	31	63.64	75	79	78	0	72	80	64	53	56	59	48	41	44	72	71	65	58	85	60	67	74	77	64	0	80	93	80	83	87	91	87	74	0	75
EXP3	29	63.98	75	79	79	0	72	81	66	51	0	54	0	50	49	72	71	65	59	84	62	69	74	79	64	0	81	92	82	83	87	92	86	74	0	76
EXP4	28	64.12	75	79	79	0	74	76	0	52	0	54	0	51	53	72	71	65	59	83	63	70	74	79	65	0	80	92	82	83	86	93	87	74	0	76

Several experiments have been carried out to obtain CI-PLUs. Discrete HMMs with four codebooks have been used in the experimentation. Moreover, in order to obtain robust acoustic models for the selected variants, the joining of sounds has been based on two criteria: less acoustic-confusion and better rates of recognition. The influence of each sound and the effect of joining sounds have been also tested. In the experiments, the phone recognition rate (%REC) has been measured as:

$$\%REC=c/(i+s+d+c)*100 \quad (1)$$

where c accounts for the number of correct recognitions, and i, s and d are the number of insertions, substitutions and deletions respectively. The following experiments have been carried out using TRAIN as training set and TUN as test:

- EXP1: with the original set of 34 units. The obtained recognition rate (61.85%) will be used as reference (table 3). The obtained result shows a clear confusion between the sounds [tsa]-[tS], [N]-[n] and [i]-[j] (table 3). Thus, these couples of sounds will be joined in the following experiments.
- EXP2: the sounds [tsa]-[tS], [N]-[n] and [i]-[j] were joined. Table 3 shows a better performance and a clear confusion among the plosives.
- Several experiments have been carried out to improve the performance of plosive sounds. Taking into account EXP2, only [B]+[b] and [D]+[d] were joined in EXP3. Table 3 shows a better performance than EXP2 and a confusion between sounds [u] and [w].
- EXP4: the vowels ([u] and [w]) were joined. Table 3 shows that this configuration leads to the best performance.

A set of 28 CI-PLUs has been finally chosen. Afterwards in order to validate the robustness of the CI-PLUs several experiments with test sets have been carried out. The results shows a %REC of 63.53% for TSDVI, 64.71% for TSIVD and 63.06 for TSIVI.

Detailed results (Table 4) show a poor performance for the plosives. The performance for these sounds could be improved by using CDSUs. In the following section a set of CDSUs will be chosen by using Decision Trees.

Finally in order to obtain robust models for all the dialects of the language a new experiment was carried out over the selected 28 CI-PLUs. In this experiment the test TSIVING was used.

Results show that this later set of units performed worse due to several factors:

- The speakers uttered [s_a] - [s] and [ts_a]-[ts] as the same sound.
- Furthermore in the case of speakers who had Basque as a second language they the uttered [S] as [s-a].
- Differences appear also in the laterals. This could be due to some rules of the automatic transcription. For example: in guipuzcoan dialect *il* is uttered as [iL] but in other dialects and specially if speakers had Basque as second language this is not always observed.

Table 4. Recognition Rates over all the sets of test with CI-PLUs.

	%REC	a	e	i	j	o	u	w	b	B	d	D	g	G	p	t	k	y	L	l	r	R	m	n	N	J	S	A	s	f	x	tsa	tt	tS	ts
TSDVI	63.53	82	79	80	0	72	74	0	49	0	45	0	50	45	78	71	74	61	84	64	68	75	80	61	0	80	63	87	81	75	87	89	84	0	77
TSIVD	64.71	76	79	79	0	74	76	0	49	0	50	0	46	60	77	71	72	72	74	64	69	78	80	65	0	83	91	86	84	88	93	90	83	0	79
TSIVI	63.06	78	79	77	0	72	77	0	47	0	43	0	53	46	78	69	73	71	68	60	69	76	72	61	0	73	75	90	85	81	91	90	84	0	78
TSIVING	55.91	79	78	75	0	68	72	0	48	0	53	0	45	50	62	57	71	51	52	51	64	76	36	56	0	47	18	74	35	68	87	44	79	0	53

5. SELECTION OF DT-CDSUs

The selection of CDSUs can improve the system efficiency by exploiting the benefits of context modelling. Decision Trees (DT) are one of the most common approaches to the selection of a suitable set of CDSUs. There are several methodologies based on the original scheme proposed by Bahl[3]. In this section several approaches have been tested.

· THE BASELINE METHODOLOGY.

The basic methodology builds a DT for each CI-PLU as follows[3]: all the samples *Y* of a CI-PLU are assigned to the root node. Then a set of binary question related to the context is used to split the node. The best question is evaluated according to the quality of the so called *Goodness of Split* (GOS) function (2), reflecting how much the likelihood of the samples increases with the split. The likelihood is calculated by Poisson discrete models. Thus *Y* is divided in two subset *Yl* and *Yr*. The process stops by two thresholds: one based on the minimum number of samples and the other one based on a minimum value of the GOS function.

$$GOS=\log\{P(Yl|Ml)P(Yr|Mr)/P(Y|M)\} \qquad (2)$$

· THE GROWING AND PRUNING METHODOLOGY.

The use of two thresholds to stop growning of the DTs has an inconvenient: for each training database some preliminary experimentation has to be carried out to obtain the optimum DT. An alternative methodology was designed to overcome this problem, the Growing and Pruning (G&P) algorithm [4]. This divides the set of training samples in two subsets: the first one to grow the DT and the second one to prune the nodes of the DT by using a misclassification measure. The algorithm converges in a few steps.

A final experiment has been carried out based on the selected 28 CI-PLUs. G&P methodology and the GOS function (2) previously defined have been integrated[5]. The TRAIN set was used to built the DT and TSIVI was used as test set. Discrete HMMs with four codebooks were used in the experimentation. A recognition rate of 68.38% was obtained with these units. This shows a clear improvement of the performance with regard to CI-PLUs (63.06%).

6. CONCLUDING REMARKS

In this work the selection of suitable sets of Sublexical Units for Continuous Speech Recognition of Basque has been carried out. Only one of the dialects is used to choose the preliminary set of sounds. Then, Context Independent Phone-Like Units are selected. The obtained units have been evaluated with most of the dialectal variants of Basque. Finally by applying the Decision Tree methodology and the G&P algorithm Context Dependent Sublexical Units have been selected. Results show a better performance of DT-CDSUs. In future works a wide analysis of these units will be made.

7. ACKNOWLEDGEMENTS

The authors would like to thank to all the people who collaborated in the manual transcription of the database and with helpful suggestions about the development. Thanks also to the anonymous Basque speakers who lent themselves unselfishly to record the database.

8. REFERENCES

[1]Mitxelena K. 1977, *"Fonética histórica vasca"*. Julio de Urquijo. Imprenta de la Diputación de Guipúzcoa, Donostia, 1977.

[2]López de Ipiña, M.K., Torres M.I., Oñederra M.L., 1995, *"Design of a phonetic corpus for Automatic Speech Recognition in Basque Language"*. Proc. EUROSPEECH'95, vol 2, pp.851-854, Madrid, 1995.

[3]Bahl L.R., Souza de V.P., Gopalakrishnan P.S., Nahamoo, D., Picheny M.A., *"Decision Trees for Phonological Rules in Continuous Speech Recognition"*, Proc. IEEE ICASSP'94, pp.533-536.

[4]Gelfand S.B., Ravishankar C.S., Delp E.J., *"An Iterative Growing and Pruning Algorithm for Classification Tree Design"*. IEEE Trans. On PAMI, vol. 13, No.2, pp. 163-174. 1991.

[5]López de Ipiña M.K., Varona A., Torres I., Rodríguez L.J., *"Decision Trees-Based Context Dependent Sublexical Units for Spanish Continuous Speech recognition Tasks"*. Proc. SNRFAI'99, vol 1, pp. 53-58, Bilbao, 1999.