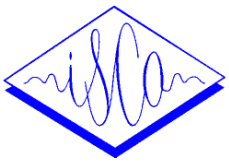

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

DESIGN OF A PHONETIC CORPUS FOR A SPEECH DATABASE IN BASQUE LANGUAGE

K. López de Ipiña[1], I.Torres[2] and L.Oñederra[3]

[1]Dpto. Automática, Electrónica e Ingeniería de Sistemas. e-mail: karmele@we.lc.ehu.es
Universidad Pública de Navarra/Nafarroako Unibertsitate Publikoa. SPAIN.
[2]Dpto. Electricidad y Electrónica. e-mail:manes@we.lc.ehu.es
[3]Dpto. Filología Vasca. e-mail:fvponoll@uc.ehu.es
Universidad del PaísVasco/Euskal Herriko Unibertsitatea. SPAIN.

ABSTRACT

The design of Continuous Speech Recognition System requires to select a large amount of spoken data for each specific language. The goal of this work was the design of a Phonetic Corpus for a Speech Database in Basque language. Several samples of nowadays narrative, spoken language and newspaper language were previously analysed under a phonetic point of view. The Speech Database finally designed consisted of a Phonetic Corpus including 300 sentences phonetically balanced uttered twice by 40 speakers resulting in about 900.000 allophones. Two additional corpora of digits and short words completed the database. This database includes the adequate distribution of allophones and contexts to model Basque phones in both, Speech Recognition Systems and Linguistic analysis frameworks.

Keywords: *Speech Databases, Basque language.*

1. INTRODUCTION.

The design of any Automatic Speech Recognition System requires a large amount of spoken data to obtain reliable acoustic models and/or adequate language models for specific tasks. Thus during last years the design of adequate Speech Databases has been an important point of interest in the Speech Recognition Community. Nowadays it is possible to find large and well-defined Phonetic Corpora and Databases for specific tasks for languages like English [12], Spanish [8] or Japanese [4]. However there is an important work to do for minority languages [3].

The goal of this work was the design of a Phonetic Database in Basque language for Automatic Speech Recognition purposes. Basque language is considered as official language, along with Spanish, for a Community of 2,5 million persons living in the Basque Country and Navarre, in the North of Spain; about 70.000 more Basque speakers can also be found in the Northeast of the Basque Country, located in the South of France. Basque is considered as one of the oldest European languages and includes some interesting typological characteristics not frequents in other European languages [1], [6], [7], [9], [10], [11]. On the other hand the use of Basque is more and more present in all the fields of nowadays Basque life: administration, industry, scientific community, mass media (radio, TV, newspaper, etc.). Thus the Basque Universities and local enterprises have increased the interest for the study of this language, not only from an academic point of view but also developing bilingual industrial applications in the speech technology framework. Any speech recognition and/or synthesis system must be bilingual in this Community since both Spanish and Basque, are official languages.

In Section 2 we deal with the phonetic analysis of the language. A complete inventory of Basque allophones is also provided. The criteria considered to design the Phonetic Corpus as well as its statistical analysis are presented in Section 3. Section 4 summarises the database characteristics and then some concluding remarks are finally presented in Section 5.

2. PHONETIC ANALYSIS OF THE LANGUAGE.

Basque Language has been poorly studied from a statistical an acoustic phonetic point of view. Thus the first goal of our work was to analyse large amount of data in order to classify the Basque phones. This language, even if spoken by a reduced Community, presents a wide dialectal distribution due to its historical isolation and old sociological and historical-political factors [7]. Since phonetic differences can be found among dialects only one of them, the *Guipuzcoan* one, was chosen for this work. This dialect presents a wide acoustic-phonetic variability [6], [7], [9], [11]. Moreover, *Guipuzcoan* is, along with Navarrese varieties, the closest to the unified language, *Euskera Batua*, which was officially instituted by the Basque Academy in 1968. A complete inventory of *Guipuzcoan* allophones is shown in Table 1 [2].

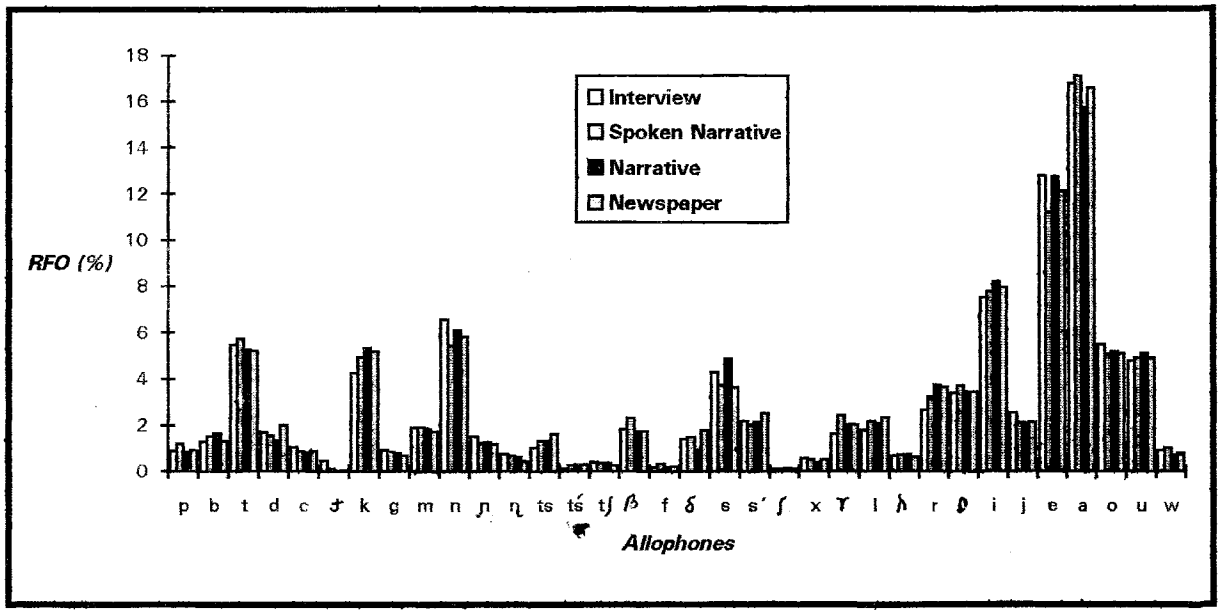

Figure 1: Relative Frequency of Occurrence (RFO) of allophones in the four samples selected as references.

1.- Training sub-corpus: 200 sentences phonetically balanced.

2.- Validation/Tuning sub-corpus: 50 sentences phonetically balanced.

3.- Test sub-corpus: 25 sentences selected from the interview reference sample and 25 sentences selected from the other reference samples.

This corpus was analysed using the interview sample as reference since there were not noticeable differences in allophone distributions among the four samples. Detailed analysis of allophones, left contexts, right contexts and left and right contexts can be found in [5]. Within the automatic acoustic-phonetic framework, this kind of information will be required in further selection of adequate sets of sublexical units. Figure 2 shows a histogram of the RFO for each allophone in both the Phonetic Corpus and the reference sample (interview sample).

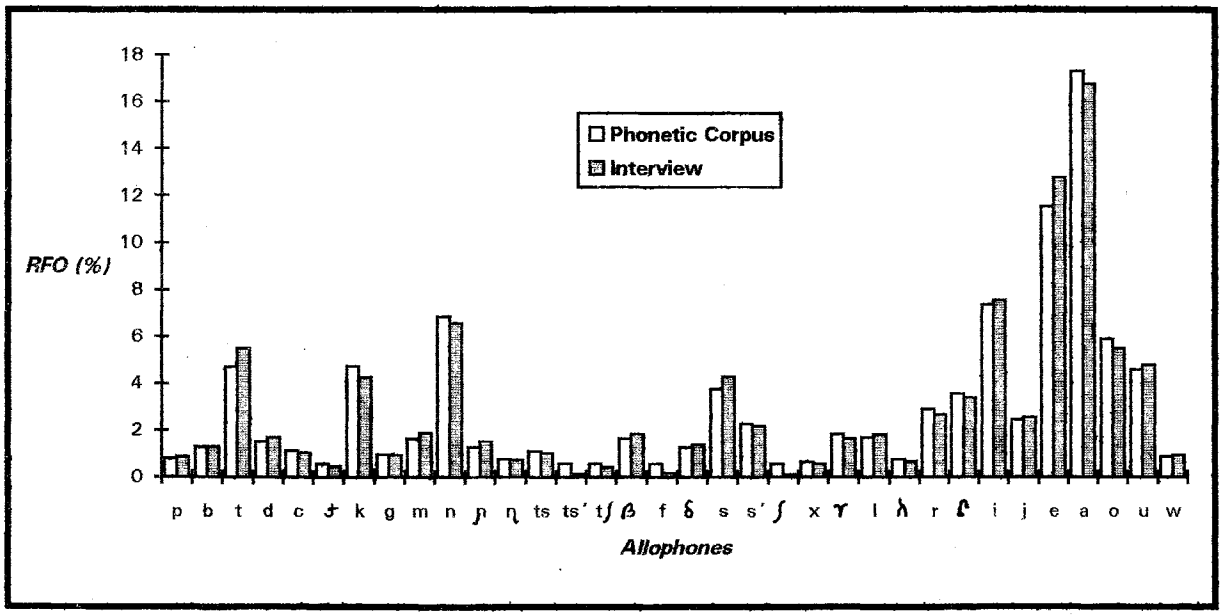

Figure 2: Relative Frequency of Occurrence (RFO) of allophones in both the Phonetic Corpus and the reference sample (interview).

These allophones are classified as:

IPA transcription

Occlusives: [p] [b] [t] [d] [c] [ɟ] [k] [g]

Nasals: [m] [n] [ɲ] [ŋ]

Africates: [ts] [ts´] [t ʃ]

Fricatives: [β] [f] [ð] [s] [s´] [ʃ] [x] [γ]

Liquids: [l] [ʎ]

Vibrant: [r] [ɾ]

Vowels: [i] [j] [e] [a] [o] [u] [w]

Table 1. Inventory of *Guipuzcoan* sounds.

IPA	SAMPA	DESCRIPTION
[p]	p	voiceless bilabial plosive
[b]	b	voiced bilabial plosive
[t]	t	voiceless dental plosive
[d]	d	voiced dental plosive
[c]	c	voiceless palatal plosive
[ɟ]	-	voiced palatal plosive
[k]	k	voiceless velar plosive
[g]	g	voiced velar plosive
[m]	m	voiced bilabial nasal
[n]	n	voiced apical alveolar nasal
[ɲ]	J	voiced palatal nasal
[ŋ]	N	voiced velar nasal
[ts]	ts	voiceless back alveolar africate
[ts´]	-	voiceless apical alveolar africate
[t ʃ]	tS	voiceless prepalatal africate
[β]	B	voiced bilabial approximant
[f]	f	voiceless labiodental fricative
[ð]	D	voiced dental approximant
[s]	s	voiceless back alveolar fricative
[s´]	-	voiceless apical alveolar fricative
[ʃ]	S	voiceless prepalatal fricative
[x]	x	voiceless velar fricative
[γ]	G	voiced velar approximant
[l]	l	voiced apical alveolar lateral
[ʎ]	L	voiced palatal lateral
[r]	rr	voiced apical alveolar vibrant trill
[ɾ]	r	voiced apical alveolar vibrant tap
[i]	i	front close vowel
[j]	j	voiced palatal approximantl
[e]	e	front mid vowel
[a]	a	central open vowel
[o]	o	back mid rounded vowel
[u]	u	back close rounded vowel
[w]	w	voiced labial-velar approximant

In order to achieve the phonetic study we selected four samples representing the more important aspects of the language. Each of them included about 25.000 allophones and 5000 words. The first one consisted of the transcription of four interviews with old people (belonging to the selected geographical area) with a poor relation with Spanish speakers (interview sample); the second one is a random selection from nowadays Basque narrative (narrative sample); the third one is also a random selection from narrative but including transcriptions of both, narrative and spoken language (dialogue, etc.) (spoken narrative sample). Finally the fourth one includes newspaper style (newspaper sample). These samples represent in an adequate way the phonetic variability of Basque language. Figure 1 shows the distribution of the Relative Frequency of Occurrence (RFO) of the allophones presented en Table 1 over each of the four samples selected. RFO distribution is quite similar in all the samples, even if they correspond to very different language contexts.

3. DESIGN AND ANALYSIS OF THE PHONETIC CORPUS.

Our second goal was to design the Speech Database. The corpus consisted of 300 sentences that were selected according to the following criteria:

- The phonetic balance was guaranteed, thus the average occurrence of each allophone in the corpus is the same as in the previous reference samples. The normalised error and restrictions defined in other studies [8] were also considered in this work to assure similar phonetic distributions in the corpus and reference samples.
- A minimum number of occurrence of each allophone was considered in order to have enough samples for further stochastic modelling (4000 realisations of each allophone in the database) [8].
- A minimum number of occurrence for the most relevant context was also considered for each allophone (400 realisations of each context in the database). All possible contexts were previously analysed. Then high RFO values and/or a strong influence in the allophone realisation defined relevant contexts [8].
- Most of the sentences were selected from the interview sample and thus are natural spoken sentences. Only a few synthetic ones needed to be added in order to guarantee the previous criteria. The 300 sentences included about 12.000 allophones.

These data will be required to obtain reliable acoustic models. Thus the design procedure also considered a partition of the corpus in three sub-corpora:

4. THE DATABASE

The selection of the speakers was made considering the dialect criteria previously mentioned. 40 speakers, 20 women and 20 men, were chosen in the selected geographical area, to pronounce the corpus. Thus, our corpus consists of 300 sentences uttered twice by 40 speakers resulting in a total of 24.000 sentences, about 170.000 words and about 900.000 allophones. The test sub-corpus was also uttered by 20 speakers selected from other geographical areas, adding about 75.000 allophones to the corpus. These speakers represent all the dialect variability of the Basque language. This database includes the adequate number of samples and balanced distribution of phones and contexts required to model Basque phones in a Speech Recognition System. Further linguistic analysis could also be made over this corpus and the reference samples considered.

Two additional sub-corpora complete the database:

1.- Word sub-corpus: 1000 short words consisting of one or two syllables were uttered by 10 speakers. These 1000 words included the more relevant contexts as well and the more confusable allophones.

2.- Digit sub-corpus: ten digits uttered twice by 20 speakers from the selected geographical area and 20 speakers from other areas. In Basque the acoustic-phonetic realisation of some digits is strongly dependent of the dialect considered.

5. CONCLUDING REMARKS.

The goal of this work was the design of a Phonetic Database in Basque language for Automatic Speech Recognition purposes. Basque language has been poorly studied under statistical and acoustic-phonetic points of view. Thus, four samples were selected from nowadays Basque narrative, newspapers and some interviews. The analysis and phone classification carried out over these samples could be considered as one of the more in depth study of nowadays Basque Phonetics. The Phonetic Corpus finally designed consisted of 300 sentences uttered twice by 40 speakers resulting in about 900.000 sounds. It includes the adequate number of samples and balanced distribution of phones and contexts required to model Basque phones in a Speech Recognition System. All the data selected and classified in this work - i.e. reference samples, Phonetic Corpus and additional short word and digit corpora - will also be the basis of further acoustic-phonetic characterisation of the Basque language.

REFERENCES

[1] R.M. de Azkue, "La Morfología Vasca", Bilbao 1925.

[2] ESPRIT Projet 2589 SAM. Final report - Year Three. SAM-UCL-G004. University College London, England, 1992.

[3] Matti Karjalaien and Toomas Altosar, "An object-oriented database for speech processing.".Proc. EUROSPEECH'93, Berlin, September 1993, pp. I-183, I-186.

[4] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis", Speech Communication, vol. 9, No. 4, August 1990, pp. 357-363.

[5] K. López de Ipiña, "Diseño de una base datos de voz en Euskera, para Reconocimiento Automático del Habla", II-3/DEE research report, 1994.

[6] L. Michelena, "Fonética histórica vasca". Julio de Urquijo. Deputation of Guipúzcoa, San Sebastián, 1977.

[7] K.Mitxelena, "La lengua vasca", Leopoldo Zugaza, Durango 1977.

[8] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mariño, C. Nadeu, "ALBAYCIN Speech database: design of the Phonetic Corpus.", Proc. EUROSPEECH'93, Berlin, September 1993, pp. I-175, I-178.

[9] M. L. Oñederra, "Euskal Fonologia: Palatalizazioa. Asimilazioa eta hots sinbolismoa", Lingua, University of the Basque Country, 1990.

[10] H. Schuchardt, "Primitiae linguae Vasconum (Einführung ins Baskische).", Halle 1923.

[11] J.L. Alvarez, "Fonologia", Ediciones Vascas, San Sebastian, 1980.

[12] V. Zue, S. Seneff, J. Glass, "Speech database development at MIT: TIMIT and beyond.", Speech Communication, vol. 9, No. 4, August 1990, pp. 351-356.