A SPEECH DATABASE IN BASQUE LANGUAGE(*)

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

[1] Elektrika eta Elektronika Saila. Zientzi Fakultat. Euskal Herriko Unibertsitatea. PK. 644 – 48080 Bilbo (Spain)

[2] Euskal Filologi Saila. Euskal Herriko Unibertsitatea. Gasteiz (Spain)

Abstract

The goal of this work was the design of a Phonetic Corpus to get a Phonetic Speech Database in Basque Language. A previous analysis of the phonetic distribution of the Language was carried out over a large amount of data in different language contexts. Then a large set of speakers from different dialectal environment was selected to record the Speech Database. The Final Speech Database included 26.000 sentences uttered by 60 speakers resulting in a total of 950.000 sounds. Further use of this material includes the development of Automatic Speech Recognition Systems in Basque Language and detailed phonetic and phonological characterisation of the language. Finally, two additional corpora consisting of sets of confusable words and digits were also designed and recorded.

1. Introduction.

Automatic Speech Recognition Systems are based on robust modelisation of adequate sets of sublexical units. To obtain such a models large amount of data are usually required. Nowadays it is possible to find large and well-defined Phonetic Corpora and Databases for specific tasks for languages as English [1], Spanish [2] or Japanese [3]. However, European minority languages require the Scientific Community to made great efforts to get such an objective [4] [5] [6]. In some cases, like the Basque language, previous phonetics analysis need also to be carried out since these languages have been poorly studied from an acoustic phonetic point of view [6] [7] [8] [9] [10].

The main goal of this work was the design of a Speech Database in Basque Language. This Database will be used not only for Speech Recognition purposes but also for phonetic and phonological analysis.

In Sections 2 and 3, the Basque Language is briefly described. A complete inventory of Basque sounds is also provided. The frequency of occurrence distribution of these sounds was experimentally obtained over four selected samples including very different linguistic contexts. Section 4 deals with the design criteria of the Corpora and Section 5 summarises the final composition of the Database. Finally, some concluding remarks are presented in Section 6.

(*) Work partially supported by the Spanish CICYT under grant TIC95-0884-CO4-3.

2. Basque Language.

Basque language is considered as official language, along with Spanish, for a Community of 2.5 million persons living in the Basque Country, in the North of Spain. About 70.000 more Basque speakers can also be found in the North of the Basque Country, located in the Southwest of France. The use of this language was restricted some years ago due to sociological and political factors. However it has spectacularly increased in the last few years and the language 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 enterprises have increased the interest for the study of this language, from not only an academic point of view but also developing bilingual industrial applications in the speech technology framework. Most of those applications, like speech recognition and/or synthesis systems, nowadays working in Spanish, or even in English, need now to be adapted to Basque language to be considered in a Community where both Spanish and Basque, are official languages.

Moreover, Basque is considered as one of the oldest European languages and includes some interesting typological characteristics that scarcely appear in other European languages [7] [8] [9] [10] [11] [12]. Some examples are its particular morphological structure and the wide variety of fricative and africate sounds. These particularities make the language more attractive to the research.

In spite of the Basque Language is spoken by a reduced community, it presents a wide dialectal distribution due to its historical isolation and old sociological and historical factors [8]. In fact, there are eight dialectal varieties [5]. This variety involves not only typological differences but, in some cases, differences in the set of sounds used by the speakers. Under a phonetic point of view, the main differences can be found in the set of africates used by speakers from different regions in the South of the Basque Country. On the other hand, French phonetics has influenced the pronunciation of Basque in the North that distinguishes it from dialects spoken in the South.

Furthermore, it is also important to consider the new unified language Euskera Batua [8] [9] [10], which was officially instituted by the Basque Academy in 1968 to overcome dialectal differences. Euskera Batua has nowadays a great importance in the Basque community since it is the language learned by adults and children at school. This unified language is nowadays the variety used by the mass media (newspapers, TV, radio, etc.) and most of Basque authors.

3. Phonetic analysis of the language.

Basque Language has been poorly studied from statistical and acoustic phonetic points of

view. Thus, the first goal of our work was to analyse large amount of data in order to classify the Basque phones and to get an inventory of sounds to be used as reference. Since phonetic differences can be found among dialects only one of them, the *Guipuzcoan* one was chosen for this work. This choice was based on two important facts: the first one is that this dialect presents the widest acoustic-phonetic variability; the second one is that the Phonetic and Phonology of the unified Euskera Batua was mainly based on this variety [8] [9] [10]. A complete inventory of *Guipuzcoan* sounds using the IPA notation is shown in Table 1.

Table 1: Inventory and detailed description 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
[J]	jj	voiced palatal plosive
[k]	k	voiceless velar plosive
[g]	g	voiced velar plosive
[m]	m	voiced bilabial nasal
[n]	n	voiced apical alveolar nasal
[ɳ]	J	voiced palatal nasal
[ŋ]	N	voiced velar nasal
[ts]	ts	voiceless back alveolar africate
[ts']	ts_a	voiceless apical alveolar africate
[t?]	tS	voiceless prepalatal africate
[β]	B	voiced bilabial approximate
[f]	f	voiceless labiodental fricative
[ð]	D	voiced dental approximate
[s]	s	voiceless back alveolar fricative
[s']	s_a	voiceless apical alveolar fricative
[?]	S	voiceless prepalatal fricative
[x]	x	voiceless velar fricative
[ɣ]	G	voiced velar approximate
[l]	l	voiced apical alveolar lateral
[ʎ]	L	voiced palatal lateral
[r]	rr	voiced apical alveolar vibrant (trill)
[*r*]	r	voiced apical alveolar vibrant (tap)
[j]	j	voiced palatal approximate vowel
[i]	i	front close vowel
[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 velar approximate

Additionally the SAMPA notation is also included. This notation was established in a European project framework [13] in order to define an ASCII symbol to represent any sound appearing in European Languages. However, Basque language includes two sounds not totally

defined in SAMPA notation: [ts'] and [s']. These sounds appear with the symbol "_a" meaning apicality but they are really apicals.

In order to get the frequency of occurrence of each sound in the language four samples representing the more important aspects of the language were selected to be analysed. Each of them included about 25.000 sounds and 5.000 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). Interviewers tried that speaker's interventions were as long and spontaneous as possible. The persons interviewed usually told some reminiscences of their life some years ago. The sample is really interesting since it constitutes a good collection of spontaneous spoken language as it used to be.
- The second one is a random selection from nowadays Basque narrative (narrative sample). The texts were selected among writers from the selected geographic environment.
- The third one is also a random selection from narrative but including transcriptions of spoken language (spoken narrative sample). The texts were selected among dialogues, songs, etc. included in nowadays novels. The language context can be considered in the limit between spoken language and narrative.
- The fourth sample includes newspaper style (newspaper sample). A set of interviews to persons living in the selected geographic environment was extracted from local newspapers.

These samples represent in an adequate way the phonetic variability of Basque Language. All the samples were analysed to get the frequency of occurrence of each sound in each sample. Figure 1 shows these distributions for the set of sounds represented in Table 1. A more detailed analysis including frequencies of right and left contexts, frequency of triphonemes, etc. was also carried out and it can be found in [5].

Figure 1 shows that the four samples present similar distributions, even if they correspond to very different language contexts. The most frequent sounds are [e] and [a] vowels. Among plosives, [t] and [k] were the most frequents, really important in Basque language. It also important to keep in mind the distribution of the africate and fricative groups because of the special features described bellow. In the first one the maximum was situated clearly in [s] and the minimum in sound [f] , being the global minimum,.

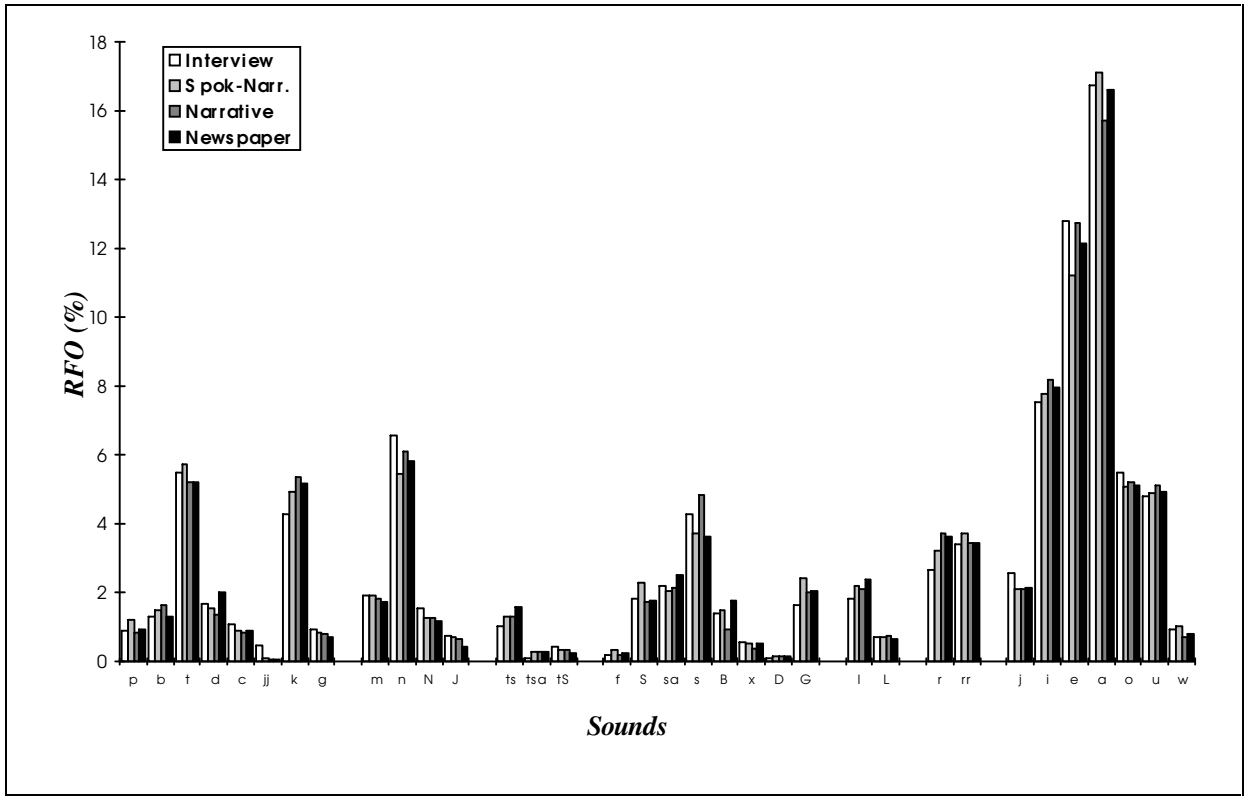

Figure 1. Relative Frequency of Occurrence (RFO) of sounds in the four samples selected as references.

4. Design and Analysis of the Corpora.

The aim of this work was to design and record a Phonetic Speech Database in Basque Language. The main goal was to obtain large amount of data to train robust acoustic models to be used in Automatic Speech Recognition Systems. However further acoustic and phonetic studies of the language will also be carried out with this material. Thus, both objectives were considered in the design of the Database that finally consisted of three corpora:

Phonetic corpus: training, tuning and testing Automatic Speech Recognition Systems.

This Corpus was designed to be used in the Automatic Speech Recognition framework. Consequently, the selection of the sentences was carried 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.
- A minimum number of occurrence of each allophone was considered in order to have enough samples for further stochastic modelling.
- A minimum number of occurrence of the most relevant context (left and right contexts)

had also to be guaranteed.

A set of 150 sentences was selected from the interview sample and thus they were natural spoken sentences. Only 50 synthetic ones needed to be added in order to guarantee the minimum number of occurrence for all the sounds and contexts. The 200 sentences included about 12.000 sounds.

This corpus was analysed using the spoken language sample as reference since there were not noticeable differences in allophone distributions among the four samples described in previous Section. Figure 2 shows an histogram of the RFO for each allophone in both the Phonetic Corpus and the Reference Sample (interview sample). Detailed analysis of sounds, left contexts, right contexts and left and right contexts can be found in [5].

This corpus constitutes the training corpus, that is, the corpus to be used to train robust acoustic models for each sound. Additionally two auxiliary corpora need also to be designed to be used when building Automatic Speech Recognition Systems: tuning and test corpora.

The Tuning Corpus is used once the basic set of acoustic models is trained to tuning the system. As a consequence the selection of sentences should also be made according to the design criteria mentioned bellow. It consisted of 25 sentences extracted from the interview sample and 25 more synthetic sentences.

Finally, the Test Corpus will be used to evaluate the Acoustic Phonetic Decoder of the System obtaining the recognition rates of the acoustic models. In this case, the phonetic balance did not need to be guaranteed. The Corpus consisted of 25 sentences extracted from the interview sample and 25 more sentences randomly extracted from the other samples. Therefore, this Corpus did not include any synthetic sentence.

Word Corpus

This corpus consisted of 1000 short words of one or two syllables. This set of words included the most confusable sounds uttered between the more relevant contexts. The object of this corpus is to obtain acoustic patterns of all the sounds of the Basque Language and to analyse which are the real differences among some specific sounds that nowadays speakers make.

Digit Corpus

Finally, the ten digits were also considered for this additional corpus. The object of this corpus is to train robust models for simple but very useful Speech Recognition applications.

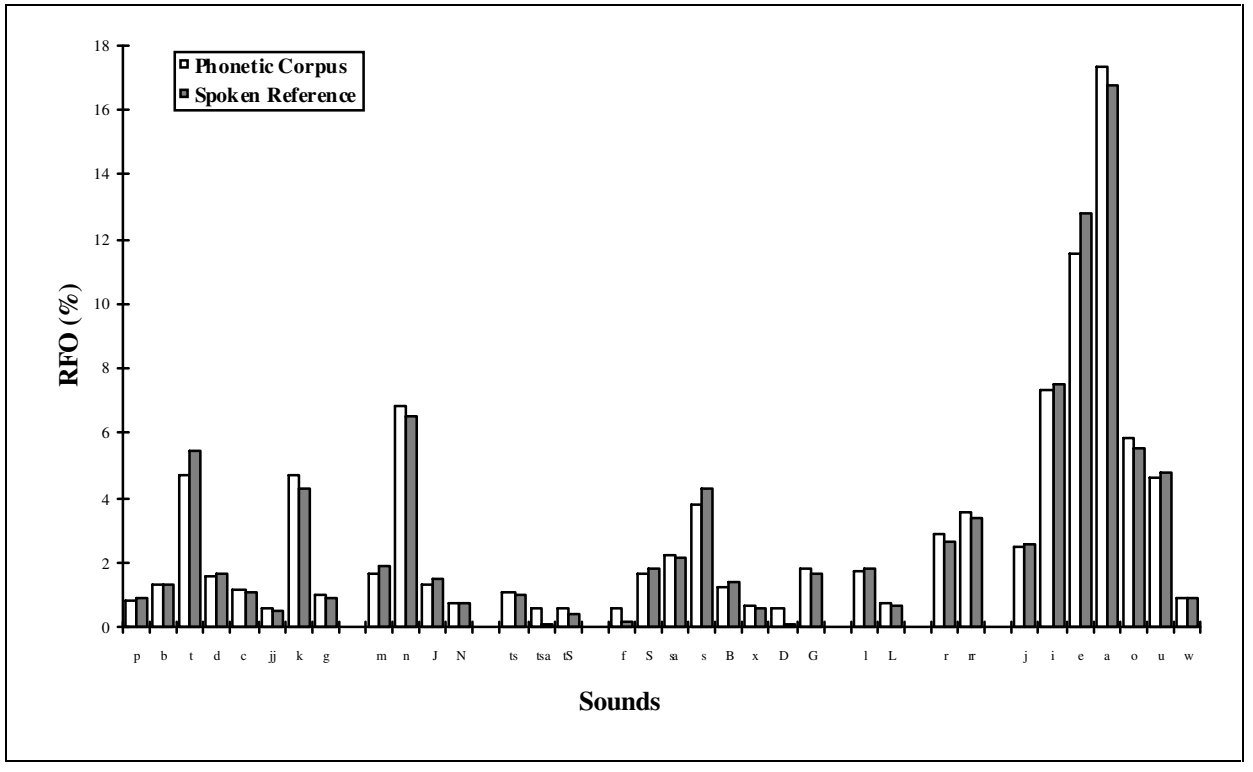

Figure 2. Relative Frequency of Occurrence (RFO) of sounds in both

5. The Database

The Database consisted of the three Corpora described in the previous Section uttered one or twice by a large set of speakers. A set of 60 speakers was selected to record the Database: 40 from the selected geographical area (Guipuzcoan ones), 10 from areas where other dialects are spoken and 10 speakers whose mother language was not the Basque (new Basque speakers). The whole Database consisted, then, of three Databases:

Phonetic Database

The Phonetic Database was based on the Phonetic Corpus described in the previous Section. The whole Phonetic Corpus was uttered twice by 40 Guipuzcoan speakers, 20 women and 20 men, resulting in a total of 24.000 sentences, about 170.000 words and 900.000 sounds. Additionally the Test Corpus (50 sentences) was also uttered twice by the 20 additional speakers. These speakers represent all the main dialect variability of the Basque language (South Country). 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. Table 2 shows the composition of this Database.

Table 2. Composition of the Phonetic Database

	Number of sentences	Number of utterances	Guipuzcoan Speakers	Other dialect Speakers	New Basque Speakers	Number of uttered sentences
Training Corpus	200	2	40			16.000
Tuning Corpus	50	2	40			4.000
Test Corpus	50	2	40	10	10	6.000

Total number of sounds: 950.000

Word Database

This database consisted the Word Corpus (1000 words) uttered by 30 speakers: 10 Guipuzcoans, 10 speaking other dialects and 10 new Basque speakers, resulting in a total of 30.000 words.

Digit Database

Finally five series of the ten digits were uttered twice by 40 speakers: 20 Guipuzcoan, 10 from other areas and 10 new Basque speakers. Let us notice that one of the differences among dialects is the pronunciation of some of the digits like zero, seven, eight and nine: sero or s_aero, saspi or s_aas_api, sortsi or s_aorts_ai and bederatsi or bederats_ai respectively.

Table 3 shows the Composition of the Word and Digit Databases.

Table 3. Composition of the Word and Digit Databases

	Number of words	Number of utterances	Guipuzcoan Speakers	Other dialect Speakers	New Basque Speakers	Number of uttered words
WORDS	1000	1	10	10	10	30.000
DIGITS	10×5	2	20	10	10	4.000

6. Concluding remarks.

The main goal of this work was the design and recording of a Phonetic Database in Basque Language for Automatic Speech Recognition purposes. However, Basque Language has been poorly studied under statistical and acoustic-phonetic points of view. As a consequence, a previous analysis of the phonetic distribution of the language needed to be previously carried out. For this purpose, four samples were selected from nowadays Basque narrative,

newspapers and some interviews with old people. The analysis and phone classification carried out over these samples could be considered one of the more in depth study of nowadays statistical and acoustic Basque Phonetics.

In a second step, a Phonetic Database was finally designed according to some previously established design criteria. This Database consisted of 300 sentences uttered by 60 speakers resulting in a total of 26.000 sentences and 950.000 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. The main important effects of the dialect distribution of the language spoken in the South of the Community were also considered.

Two additional corpora were also designed and recorded. The first one consisted of a large set of short words including most confusable sounds and designed to help further acoustic characterisation of Basque sounds. The second one consisted of the ten digits uttered by 40 speakers that will be used for simple, but robust, Speech Recognition applications.

References

[1] 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.

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

[3] 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.

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

[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] K. López de Ipiña, I. Torres and L. Oñederra, "Design of a phonetics corpus for a speech database in Basque", Proc. EUROSPEECH'95, Madrid, September 1995, pp. II-851, II-854.

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

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

[9] M. L. Oñederra, "Euskal Fonologia: Palatalizazioa. Asimilazioa eta hots sinbolismoa", Colleccion Lingua. Servicio Editorial de la Universidad del País Vasco, 1990.

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

[11] L. Michelena, "Fonética histórica vasca". Julio de Urquijo. San Sebastián: Imprenta de la Diputación de Guipúzcoa, 1977.

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

[13] Esprit Project 2589 (SAM). "Multi-lingual Speech input/output assessment, methodology and standardisation". February, 1991.