

MORENO, A.- POCH, D.- BONAFONTE, A.- LLEIDA, E.-
LLISTERRI, J.- MARIÑO, J. B.- NADEU, C. (1993)
"ALBAYZÍN Speech Database: Design of the Phonetic
Corpus", in *Eurospeech'93. 3rd European Conference on
Speech Communication and Technology*. Berlin, Germany, 21-
23 September 1993. Vol. 1, pp. 175-178. ISSN: 1018-4074.

[http://liceu.uab.es/~joaquim/publicacions/
Moreno_et_al_93_Albayzin_Phonetic_Corpus.p
df](http://liceu.uab.es/~joaquim/publicacions/Moreno_et_al_93_Albayzin_Phonetic_Corpus.pdf)

ALBAYZIN SPEECH DATABASE: DESIGN OF THE PHONETIC CORPUS

A. Moreno*, D. Poch**, A. Bonafonte*, E. Lleida*, J. Llisterra**, J.B. Mariño*, C. Nadeu*

*Universitat Politècnica de Catalunya, **Universitat Autònoma de Barcelona
Barcelona, Spain

ABSTRACT

This paper describes the phonetic content of Albayzin, a spoken database for Spanish designed for speech recognition purposes. A statistical study of a large sample of spontaneous speech is presented, and the phonetic and statistical criteria for the final constitution of the database are discussed. Finally, the contents of the phonetic database are analyzed.

Keywords: Spoken Databases

1. INTRODUCTION

ALBAYZIN is a Spanish spoken database designed for speech recognition purposes. The database is divided in three parts: a) a phonetic database, b) an application database, c) a Lombard speech database.

The corpus of the phonetic database is divided in two parts:

Subcorpus 1: Composed by utterances from a set of 200 phonetically balanced sentences. The meaning of "phonetically balanced" is described below. 4 talkers utter the overall set and 160 talkers a phonetically balanced subset of 25 sentences. As a result, we obtain 24 utterances of each sentence. Subcorpus 1 is intended for training purposes.

Subcorpus 2: Composed by utterances from a set of 500 phonetically balanced sentences. 40 talkers utter a group of 50 sentences. So, each sentence is recorded by 4 speakers.

This work was supported by a Grant from the Spanish Government TIC 91-1488-C06-02

Subcorpus 2 is designed to test the phonetic decoding performance of a recognition system.

The application database is composed by 3900 sentences. The selected application is information retrieval from a geographic database. The corpus is divided in a training set composed by 2700 sentences and a test set composed by 1200 sentences. The corpus has been divided in 78 subsets of 50 sentences each and is uttered by 136 speakers.

The Lombard database is composed by a subset of the above mentioned databases and is uttered by 40 speakers. During recordings, a noise source is applied to the speakers via headphones to produce the Lombard effect.

An equal number of male and female speakers has been chosen, providing a good coverage of ages from 18 to 55.

The design of the phonetic database is described in this paper.

2. STATISTICAL STUDY OF A SAMPLE OF SPANISH.

A study of the statistical properties of allophone distribution in Spanish was undertaken on a corpus of spontaneous speech recorded from three speakers. An orthographic transcription was made and was introduced to an automatic grapheme-to-allophone conversion programme [3]. More than 100.000 allophones were counted. Statistics include: Relative Frequency of Occurrence (RFO) of each allophone, RFO of predecessors and successors for each allophone, RFO for each

allophone in stressed and non stressed syllables, RFO of the position of each allophone in the syllable and in the word, word length in number of allophones and distance of each allophone from the stressed syllable in the word.

The statistical results obtained from this large Spanish sample will be used as reference in the design of the database.

a) Occurrence of Spanish allophones

The selection of allophones to be included in the database has been made in two steps: first, a complete inventory of allophones was considered [1,2]; second, those allophones with a RFO lower than 0.1% were discarded. The result is a set of 31 allophones as shown in Table I.

IPA	SAMPA	
p	p	voiceless bilabial plosive
b	b	voiced bilabial plosive
t	t	voiceless dental plosive
d	d	voiced dental plosive
k	k	voiceless velar plosive
g	g	voiced velar plosive
m	m	voiced bilabial nasal
n	n	voiced alveolar nasal
ɲ	J	voiced palatal nasal
ŋ	N	voiced velar nasal
tʃ	tS	voiceless palatal affricate
β	B	voiced bilabial approximant
f	f	voiceless labiodental fricative
θ	T	voiceless interdental fricative
ð	D	voiced dental approximant
s	s	voiceless alveolar fricative
Z	z	voiced alveolar fricative
ʝ	jj	voiced palatal fricative
x	x	voiceless velar fricative
ɣ	G	voiced velar approximant
l	l	voiced alveolar lateral
ʎ	L	voiced palatal lateral
r	r r	voiced alveolar trill
ɾ	r	voiced alveolar tap
i	i	front close vowel
j	j	voiced palatal approximant
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

TABLE I Allophone inventory

b) Context

All possible contexts for each allophone were analysed and the most relevant contexts were selected according to phonetic and statistical criteria. A relevant context was defined if its frequency of occurrence relative to the allophone under consideration is very high (>10%) or if it affects in a particular and distinctive way to the allophone under consideration. In order to obtain the most relevant contexts the following criteria were established:

- 1- Each allophone must be followed and preceded by a front vowel, a central vowel and a back vowel.
- 2- Consonant clusters and other contexts with important allophonic variations should appear: Consonant clusters: B-l, b-l, B-r, D-r, d-r, G-l, g-l, G-r, p-l. Other contexts: /-g, /-l, /-s, k-t, l-B, l-G, l-T, n-T, rr-B, T-p, z-B. (where /=pause)
- 3- Phonetic contexts have been arranged into categories according to the influence of the context on the target allophone. The following categories have been created: [f,s,n,m,t], [L,ts], [i,e], [D,m,B,n]. The following groups are defined for vocalic target allophones: [p,b,m,f,B], [t,d,n,s,l,r,rr,z,T,D], [k,g,G,J,ts,x,L,z,N]. It is assumed that the consonants within the same category will show similar coarticulatory effects on the target allophones.

c) Stressed syllables.

The occurrence of each allophone in stressed and non stressed syllables was also computed since this information is considered important for recognition purposes.

d) Position in the syllable.

Consonant allophones which occurs in syllable coda in Spanish are: [m,n,s,l,r,rr].

3. DESIGN OF THE PHONETIC CORPUS

The way in which the notion of "phonetic balance" was applied in the design of the corpus is explained in this section. The selection of the sentences in the Subcorpus 1 was made according to the following criteria:

- 1) Allophone. The following normalized error is defined to evaluate the degree of

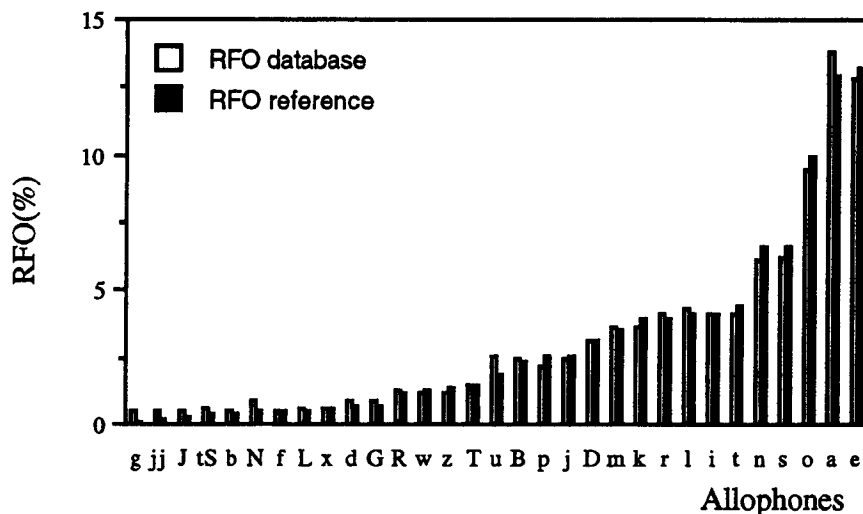


Fig.1 Histogram of Relative Frequency of Occurrence (RFO) of allophones in reference and subcorpus I

matching between the frequency of allophone occurrences in the phonetic database and in the reference.

$$E_n = \left| \frac{f(n) - \hat{f}(n)}{f(n)} \right| 100\% \quad (1)$$

Where $f(n)$ and $\hat{f}(n)$ are the RFO of allophone n in the reference and in the corpus respectively.

Sentences were selected to accomplish:

$$E_n < 15\% \text{ and } \hat{f}(n) > 0.5\% \quad \forall n.$$

The former restriction assures a phonetic distribution similar to the one found in other studies [4]. The latter restriction is imposed to be sure that all the allophones appear significantly for recognition purposes (960 realisations of each allophone in the database).

2) **Context**: A minimum number of occurrences for each relevant context was imposed as a constraint in the design. (96 realisations of each context in the database).

3) **Stressed syllables**: Each allophone occurs both in stressed and unstressed syllables with a normalised error lower than 20%.

4) **Position in the syllable**. Consonant allophones occurring in syllable coda in Spanish [m,n,s,l,r,rr] should appear significantly in this position in the corpus.

The selection of sentences for subcorpus I was made from the originally recorded database. 830 sentences were obtained by segmenting the text into sentences of two or three seconds (14-20 syllables each). Sentences were slightly modified to give semantic coherence and naturalness.

A first selection of 176 sentences was made to comply with restrictions on RFO of allophones and contexts. The remainder 24 sentences were chosen by a semiautomatic method to match the error level bounds. 9 additional sentences were manually produced to match all the specifications.

The selection of the sentences in the Subcorpus II was done according to the following allophone balancing criterium:

$$E_n < 15\% \text{ and } \hat{f}(n) > 0.25\% \quad \forall n.$$

The selection of sentences of subcorpus II was made from 3 texts of Spanish contemporaneous writers. The less represented allophone has 224 realisations.

4. ANALYSIS OF THE DATABASE

Fig1 shows an histogram of the relative number of occurrences of each allophone in the reference and in the subcorpus I. The normalized error E_n is lower than 15% in all the allophones except in [g,Z,J,ts,b,N,f,L,x] because they are affected by the additional constraint of minimum number of realisations of each allophone.

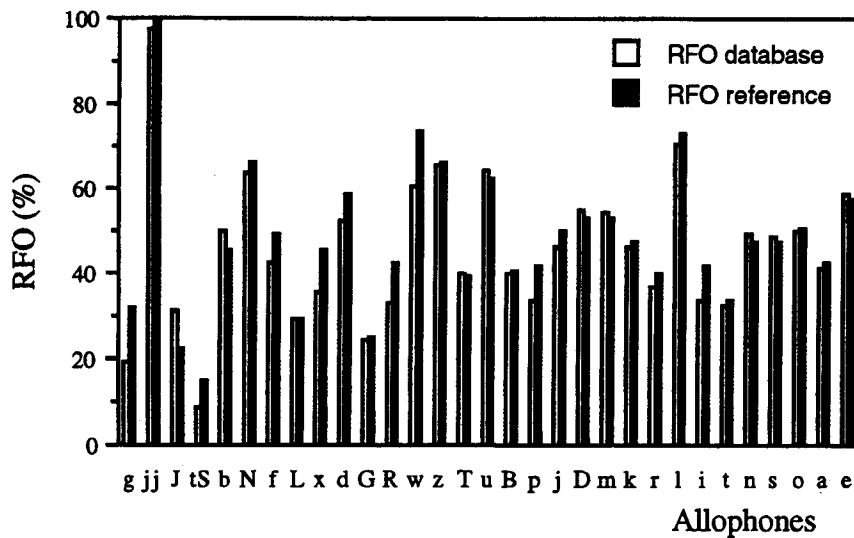
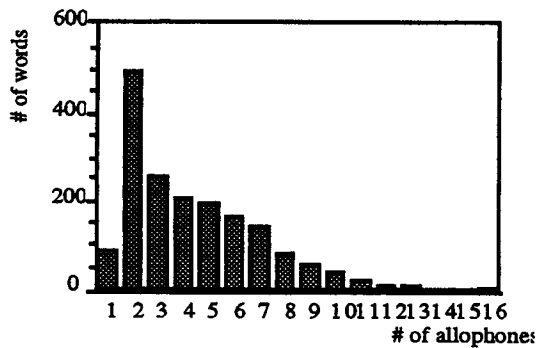


Fig. 2. Relative Frequency of Occurrence (RFO) of allophones in stressed syllable in the reference and in subcorpus I.

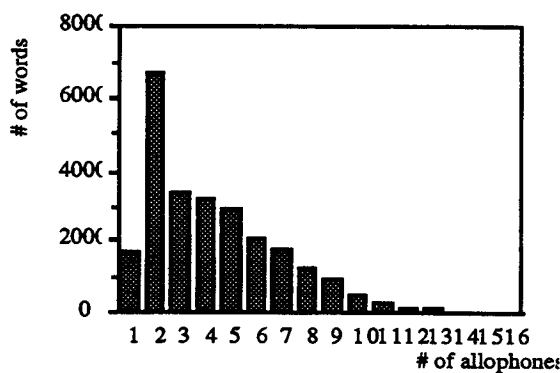
Fig. 2. shows the histogram of occurrences of each allophone in stressed syllable in the reference and in subcorpus I. The normalized error is lower than 20% for each allophone. Allophones [g,Z,J,ts,b,N,f,L,x] cannot be used for training separately stressed and unstressed occurrences

because of the low number of realisations for both cases. For this reason, [g,Z,x] were not forced to match that the normalized error in stressed and non stressed syllable is lower than 20%.

Fig 3. shows the histogram of the number of allophones per word. The distribution is very similar in the corpus a) and in the reference b)



a)



b)

Fig 3. Histogram of number of allophones per word a) in the subcorpus I, b) in the reference.

5 CONCLUSIONS

The design of a Spanish phonetic database has been discussed in this paper. A large sample of Spanish spontaneous speech has been statistically analyzed in order to define the frequency of occurrence of allophones and their contextual distribution. The results have been applied to define the phonetic content of a spoken database. Design criteria and results have been presented

REFERENCES

- [1] Navarro Tomás, T. (1918) *Manual de pronunciación española*. Madrid: CSIC 21. edición, 1982
- [2] Canellada, M.J., Kuhlman Madsen, J.(1987) *Pronunciación de español. Lengua hablada y literaria*. Madrid. Castalia
- [3]. Llisterri J. , Mariño J.B. (1993) *Spanish adaptation of SAMPA and automatic phonetic transcription*. ESPRIT project 6819 (SAM-A) Internal report.
- [4] Rojo,G. (1991) *Frecuencia de fonemas en español actual*. Ed. Brea, M. Fernández Rei, F. *Homenaxe ó Profesor Constantino García*. Univ. Santiago de Compostela. Vol I pp 451-467