A CONTINUOUS WORD RECOGNITION SYSTEM

based on an

ACOUSTIC - PHONETIC APPROACH:

an application to a small vocabulary

in ITALIAN

M. G. DI BENEDETTO
P. MANDARINI

description of the vocabulary

- Ten digits

+ 7 control words

digits	control words
uno	comando
due	invio
tre	aiuto
quattro	annullo
cinque	precedente
sei	successivo
sette	confermo
otto	
nove	
zero	

'uno

d'ue

tr'e

kw'at:ro

tʃ'iŋkwe

s'eˠ

s'ɛt:e

'ɔt:o

n'ɔve

dz'ɛro

kom'ando

iŋv'io

a'juto

an:'ul:o

pretʃed'ɛnte

sutʃ:es:ivo

komf'ermo

description of the vocabulary

phonemes present in the vocabulary

- vowels : [i, e, ɛ, a, o, ɔ, u]

  all vowels of the Italian vowel system

- stops : present [d] [p, t, k]

  miss [b, g]

- affricates : present [tʃ, dz]

  miss [ts, dʒ]

- fricatives : present [f, s] [v]

  miss [z]

- liquids : present [l, r]

  miss [ʎ]

- nasals : present [n, m]

  miss [ɲ]

  + three allophones of [n] : [*n , ŋ , ɱ]

  *n : before dentals [t, d]

  ŋ : before velars [k, g]

  ɱ : before labiodentals f, v

- semiconsonants : present [j, w]

  miss NONE

description of the vocabulary

- stops [d] word initial position
  [p] in cluster pr
  [t·] geminate intervocalic position
  " in cluster tr
  [t] in cluster tr
- affricates [tʃ] word initial position
  intervocalic position
  geminate in intervocalic position
- fricatives [s] word initial position
  geminate in intervocalic position
  [f] after ŋ
  [v] intervocalic position
- liquids [l] geminate intervocalic
  [r] in clusters tr pr
  intervocalic
  in group erm
- nasals [n] intervocalic, initial, geminate intervocalic
  [m] intervocalic and after [r]

description of the vocabulary

- the words can be pronounced in sequences of one to three words.

only one word	one to three words
annullo	digits
precedente	comando
successivo	invio
confermo	

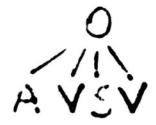
- 7 broad phonetic classes

1- Vowel-like	D
2- Vowel	V
3- Sonorant	J
4- intervocalic nasal	N
5- stop	O
6- affricate	A
7- fricative	F

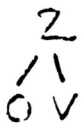
- 2a front vowel FV
- 2b back vowel BV

- vowel-like [e^Y] [w, j]
- vowel [i, e, ɛ, ə, ɔ, o, u]
  - front vowel [i, e, ɛ]
  - back vowel [ə, ɔ, o, u]
- sonorant [ɾ, l] [ŋ, m, *n]
- intervocalic nasal [n, m]
- stop [d, t, k, p]
- affricate [dz, tʃ]
- fricative [s, f] [v]

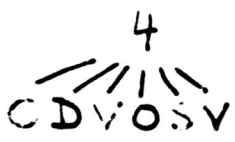
description of the [illegible] BROAD PHONETIC CATEGORIES

0


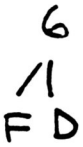
1


2


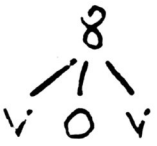
3

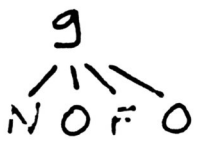

4


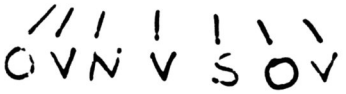
5

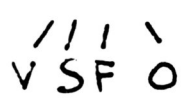

6


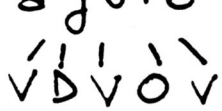
7

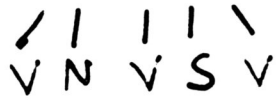

8


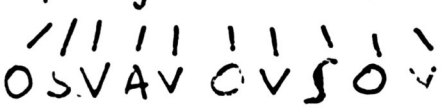
9


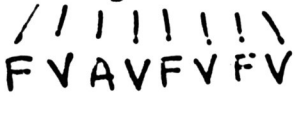
Kom'a*ndo


inv'io


a'juto


ah:'ul:o


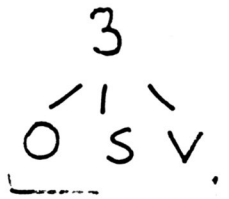
pretʃe d'ɛ*nte


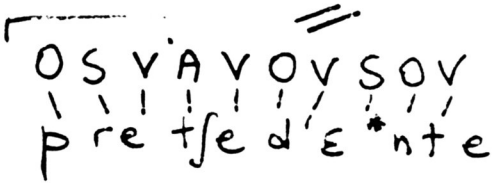
sutʃ:es:iro


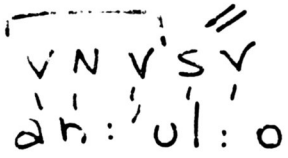
kongfermo


description of the vocabolary

PREFIX RULE NOT SATISFIED







pronounced

ONLY

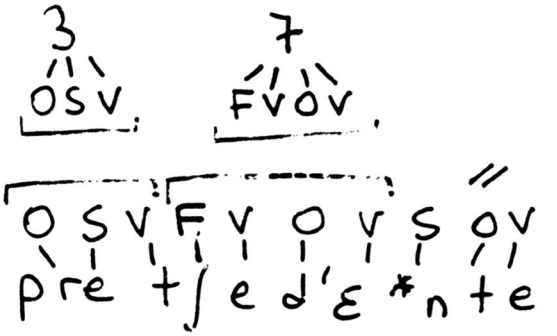
in isolation -

description of the vocabulary

differences in pronunciation

tʃ ➡ ʃ

➡ prefix rule not satisfied



↳ pronounced only in isolation -

	p	b	f	v	t	d	ts	dz	s	z	k	g	tʃ	dʒ	ʃ	m	n	ɲ	l	ʎ	r	i	e	ɛ	a	ɔ	o	u	j	w
vocalic	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	+	+	+	+	−	−
consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−
nasal	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+												
compact	−	−	−	−	−	−	−	−	−	−	+	+	+	+	+				−	+	−	−	−	−	+	−	−	−		
diffuse																						+	−	−		−	−	+		
grave	+	+	+	+	−	−	−	−	−	−	+	+	−	−	−	+	−	−				−	−	−		+	+	+	−	+
acute																	−	+												
tense																							+	−		−	+			
voiced	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−															
continuant	−	−	+	+	−	−	−	−	+	+			−		+				+	−										
strident			−	−	−	−	+	+																						

Description of the vocabulary

Distinctive Phonetic Features

- 75 sequences formed by one to three words belonging to the vocabulary

- sequences selected in order to include several types of boundaries

  Vowel - Vowel

  Vowel - Consonant

  Consonant - vowel

- ten speakers : five female and five male repeated the 75 sequences twice

- recording in a sound-proof room using high-quality equipment

- speech materials sent on the telephone channel using artificial mouth and Telephone-quality speech recorded - different telephone lines of the same telephone trunk.

experimentation

DATA BASE

- speech material digitalized : (toll quality)
  - low-pass filter 4.5 kHz
  - sampling frequency 10 kHz
  - 12 bits A/D

➡ each sentence in a file (stored on diskettes)

ILS compatible

- speech material of five speakers MANUALLY segmented using the following tools :
  - acoustic waveform
  - spectrum every 10 msec -

Voiced - Unvoiced - Silence Classifier

based on the criteria described in :

"A Bayesian - Adaptive decision method for the V/UV/S classification of segments of a speech signal"
IEEE - ASSP, vol 35, n° 4, 1987 -
by Bruno, Di Benedetto, Gilio, Mandarini -

- 4 parameters :
  - Zero - Crossing
  - Log-energy
  - 1° autocorrelation coefficient
  - 1° linear prediction coefficient

- definition of a-priori probabilities of the classes V-UV-S $P(V)$ $P(UV)$ $P(S)$

- definition of transition probabilities between classes $P(V/UV)$ $P(UV/S)$ $P(S/UV)$
  $P(UV/V)$ $P(V/S)$ $P(S/V)$
  $P(V/V)$ $P(UV/UV)$ $P(S/S)$

PRELIMINARY RESULTS

- training on two speakers - 10 ms (100 samples)

	V	UV	S
zero crossing	low ∼9	high ∼40	medium ∼20
log-energy	high ∼45	medium ∼29	low ∼19
1° autoc. coef	high 0.95	low 0.25	medium 0.65
1° LP coeff.	high 1	low -0.6	medium .1

- average values (mean vectors)

  variances (covariance matrices)

computed for each class -

experimentation | V-UV-S classifier

preliminary results

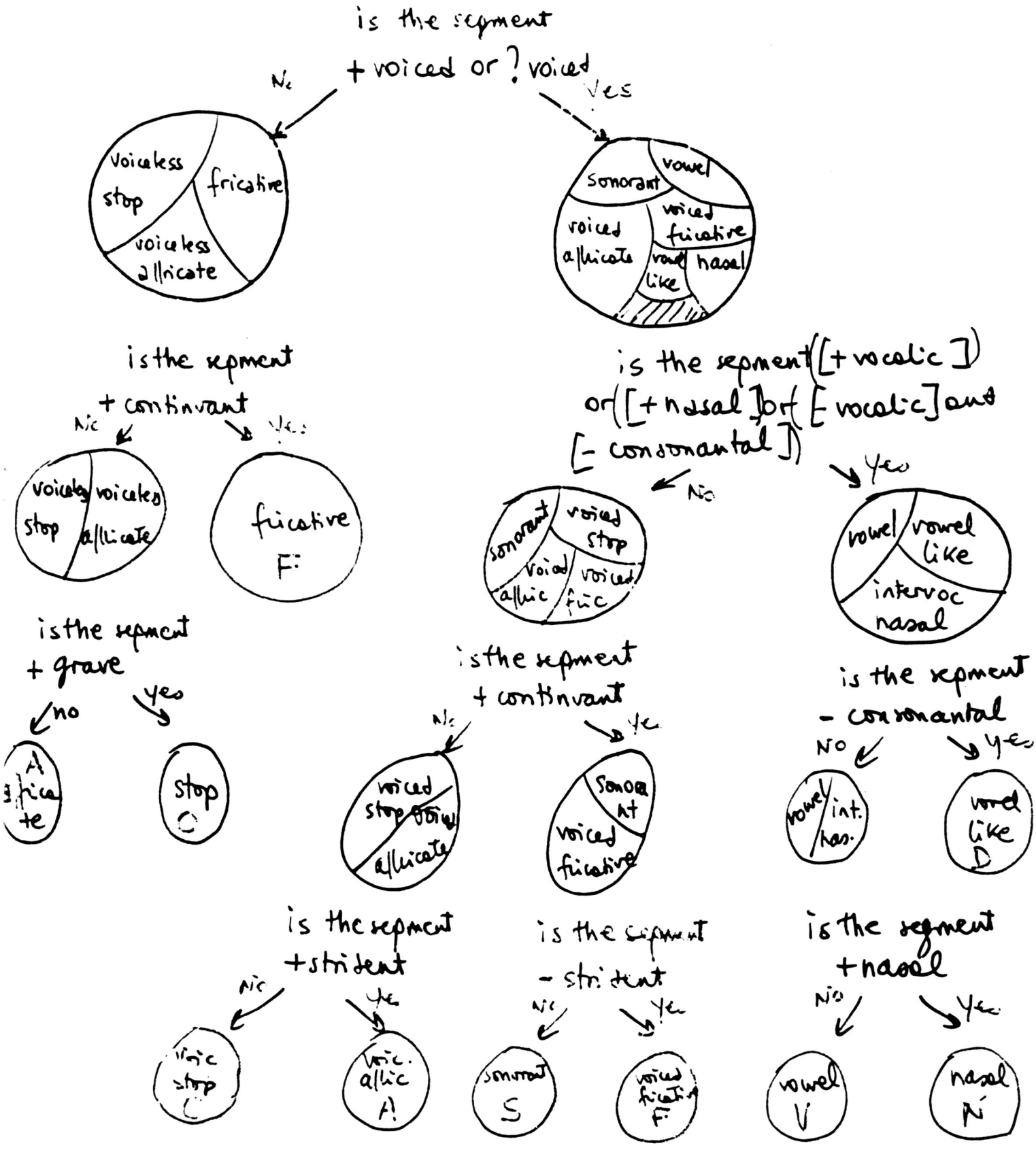
on the same speakers of the Training

in \ out	V	UV	S
V	0.89	0.07	0.04
UV	=	0.79	0.21
S	=	=	1

error types

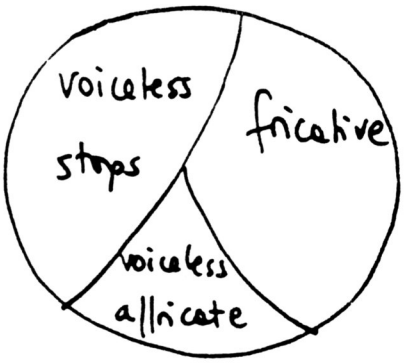
in \ out	V	UV	S	
V		[d] [v] [dz] [z]	0.04 [v]	
UV			0.14 [k] in 0.04 [f] first 0.03 [t] or	initial position frame [p] in tr/pr
S				

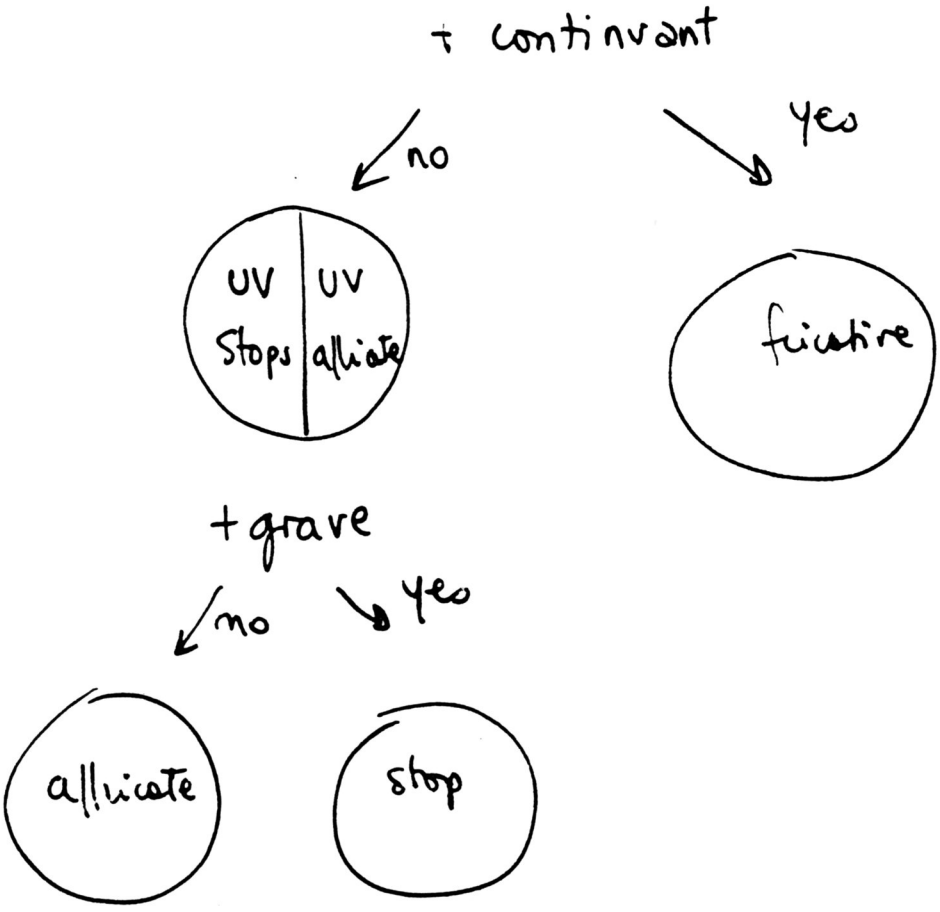
recognition strategy



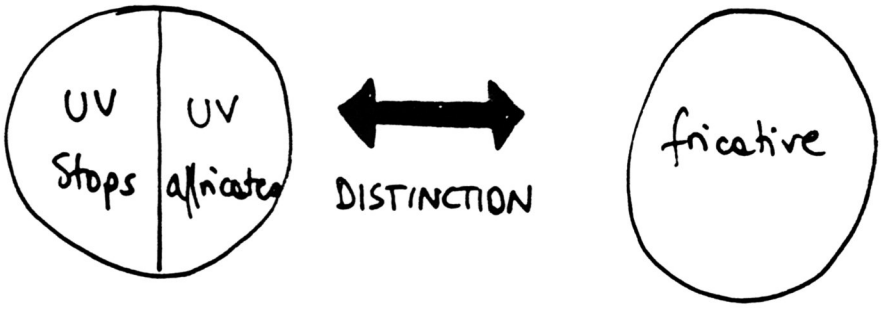
experimentation | - voiced segments

V-UV-S Classifier produces S and UV segments which are - voiced.





acoustic analysis



- sequence

  V-S-UV-V ⇒ stop or affricate

  V-UV-V ⇒ fricative

⚠ rule not valid for UV in initial sequence position.

- energy parameter

$$E_1 = \frac{\text{Energy between } 2500 \div 3300 \text{ Hz}}{\text{Energy between } 300 \div 3300 \text{ Hz}}$$

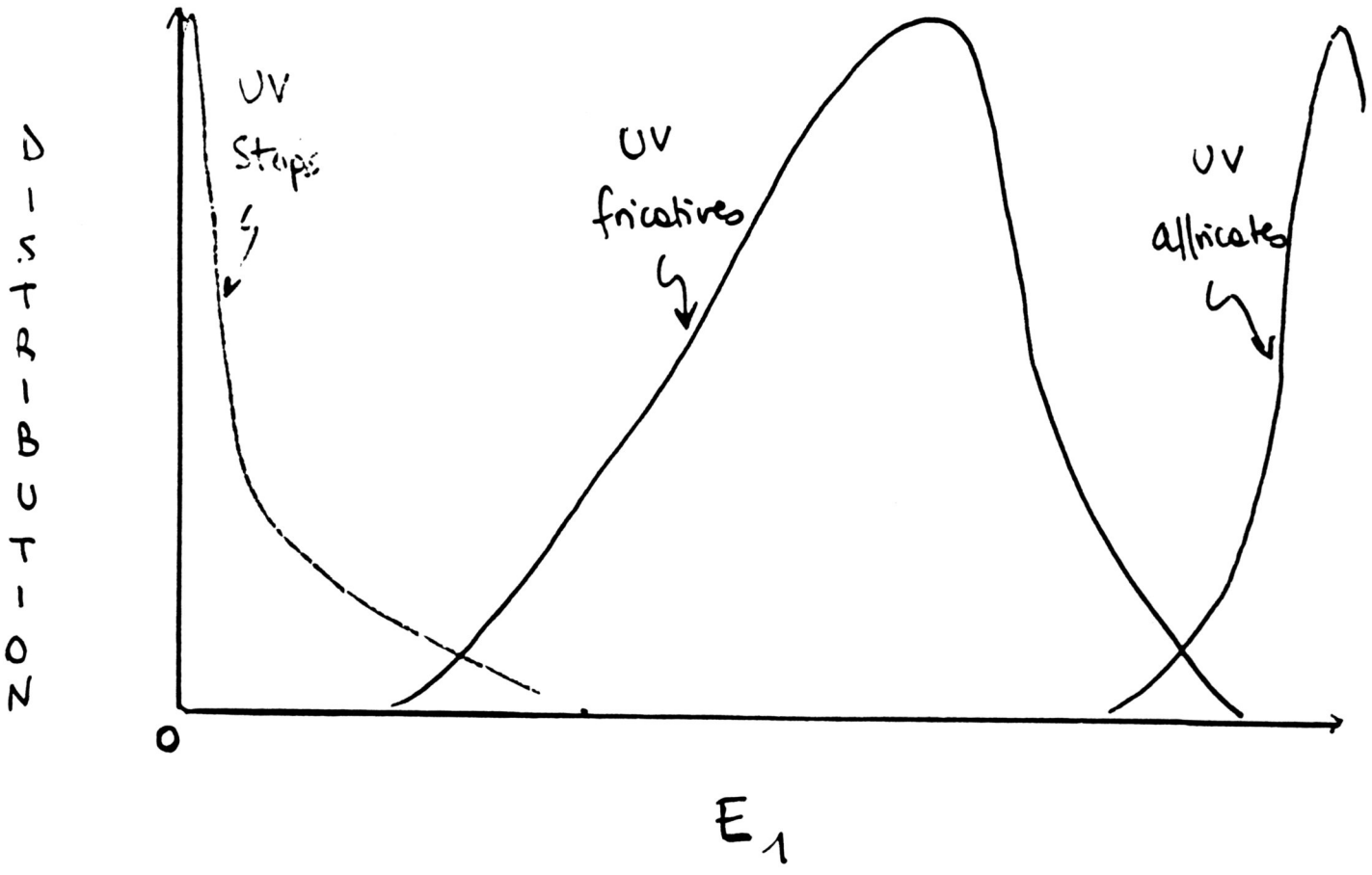
telephone channel bandwidth.

acoustic analysis

distinction + continuant (fricative)

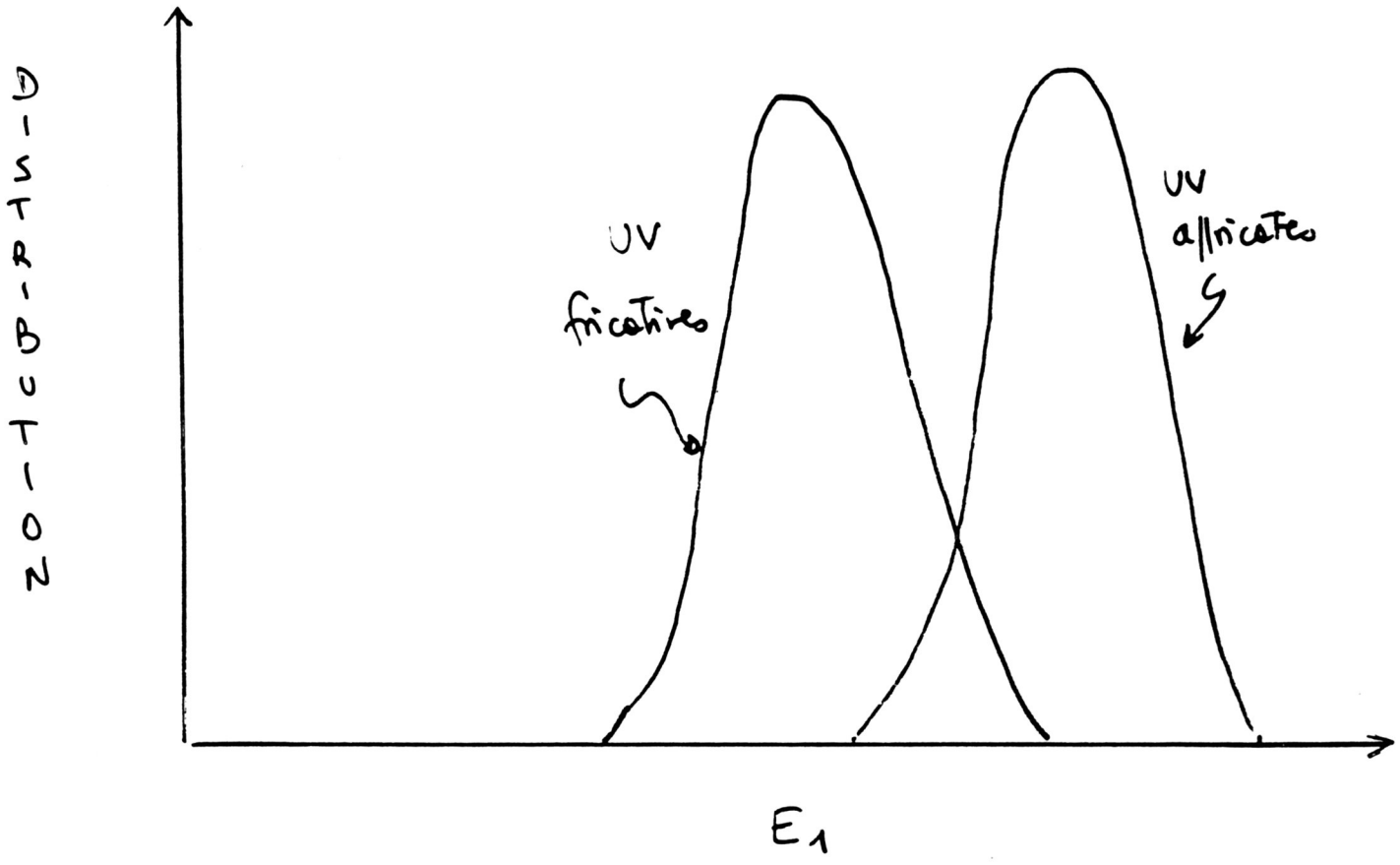
− " (affricate + stop −

parameter : E_1 = $\frac{\text{energy between } 2500 \div 3300}{\text{energy between } 300 \div 3300}$.



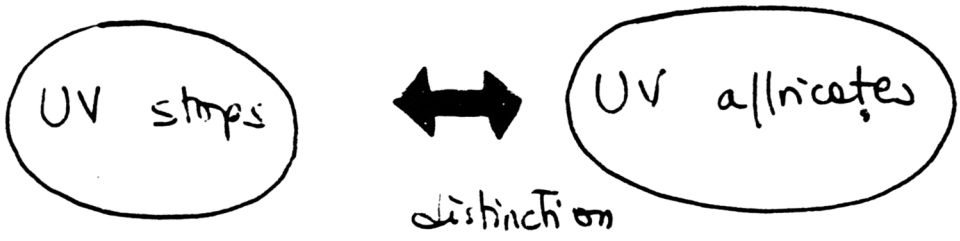
distinction affricates – fricatives

E_2 = energy between 1600 and 3300 Hz



experimentation

acoustic analysis



- E_1 as seen previously
- burst duration

UV stops average burst duration

25 msec

UV allricates average burst duration

60 msec

Conclusion

- application of the acoustic-phonetic approac to a small vocabulary in Italian connected words
- data-base existing and digitalized
- data-base segmented
- preliminary results on:
  - distinction V-UV-S
  - distinction stops-affricates-fricatives
  - distinction stops-affricates

Future Work

- extension to all classes
- " to other speakers
- " to telephone quality speech.