PHONETIC CODING FOR DATA BASES AND EXPERT SYSTEMS

GILBERT PUECH et PIERRE BANCEL

Centre de Recherches Linguistiques et Sémiologiques
Université Lumière-Lyon 2
69500 Bron, France

ABSTRACT

The International Phonetic Alphabet (IPA) is the standard reference as a transcription system. With only minor variants, it is commonly used by linguists to record the pronunciation of languages whether they are supported by an orthographic tradition or not. The scope of this paper is to transpose the IPA to a computer-oriented coding system in order to use phonetic records in data bases and expert systems.

INTRODUCTION

A computer-oriented coding system for the representation of sounds should be viewed as an interface between linguists faced with the representation of a wide range of sounds and a Data Base Management System.

First the code corresponding to each sound must be a key to its major characteristics and, consequently, to the way it patterns with other sounds. The binary distinctive features theory seems to be the natural interface between phonetic analysis and the binary logic of computers. It turns out, however, that there is no clear agreement on how a number of complex or rare sounds should be treated in this approach; furthermore the built-in definition of some features is costly since it precludes some combinations - for instance [+High] is exclusive of [+Low] - or hardly satisfactory to account for some sounds - such as flaps and trills. On the other hand an IPA based classification presents several advantages : it is independant of any particular theory ; it associates phonetic interpretation and a graphic representation in the same table ; it allows a more compact code. This code can be easily converted into a matrix of distinctive features so that the exploitation of the data can be independent of the coding system.

Secondly, the coding system must fit one of the standard formats for computer words. It should also be used to classify phonetically recorded words in the data bases in the same manner as the ASCII code is used to classify orthographically recorded words. If the data base is organized in n-ary trees, the algorithm will find all the relevant information necessary for the equilibration of the trees in the set of codes forming each word.

GENERAL ORGANIZATION

For maximal efficiency, each segment is coded in a short integer (16 bits word) noted by 4 hexadecimal figures. Consonants and vowels are coded independantly of each other, thus it is necessary to know if one given code refers to a consonant or to a vowel before being interpreted. For languages - such as Bantu - in which words are built after a strict syllabic pattern, the data base may determine the fields composing the records as corresponding either to a consonant or to a vowel; in languages where no such syllabic regularity prevails, the first field of the record (a long integer) will in the first byte determine the number of segments included in the record and, in the three following bytes, select the V/C choice (bits 8 up to 31 set to 1 when the segment should be interpreted as a vowel and left at 0 if it is a consonant). Suprasegmental information - stress and pitch - is normally associated with vowels; provision is made however for consonants bearing a tone. A set of diacritics is used to give maximal versatility to this coding system which was designed both for narrow and broad transcriptions. Coding of morpheme boundaries for morphophonemic representations was not examined but could be accomodated.

CONSONANTS

A - Basic consonants are coded in the least significant byte of the short integer. Table 1 yields the phonetic interpretation of the coding and illustrates some of the realizations. The 4 most significant bits correspond to the lines (manner of articulation) and the 4 remaining bits to the columns (place of articulation) :

Phonetic symbol	Code	Phonetic interpretation
b	0041	bilabial voiced stop
m	00C1	bilabial nasal stop
kp	001C	labiovelar unvoiced stop

Sonorants (lines B to F) are assumed to be voiced; implosives and ejectives are respectively voiced and unvoiced. For clicks, which may be voiced, aspirated, murmured etc., further qualification is needed. In order not to have more than 15 places of articulation, some choices had to be made; thus, apico-labial sounds, which are to be found in Umotina [1], are not included in the set of basic consonants but could be handled as a special case (see section F). To facilitate the editing on the lineprinter, it is convenient to have each basic symbol occupy one space only even if it is commonly transcribed as a sequence of two consonants (such as kp or ts).

B - Double consonants, geminates as well as complex segments, are coded in two morae and occupy two spaces:

bb	4141	geminate bilabial voiced stop
mb	C141	bilabial prenasalized voiced stop
nt	C414	alveolar prenasal. unvoiced stop
nts	C474	alveolar prenasalized unvoiced affricate

C - A release, transcribed by a right-adjacent diacritic occupying half a space, is coded in the least significant byte: the most significant bits refer to Table 2; the final hexadecimal zero is a flag indicating that the basic consonant (coded in the first byte) is followed by a release, the interpretation of which is given in Table 2:

kʸ	1B90	velar stop/palatal release
bʷ	41C0	bilabial stop/labiovelar release
dʳ	44F0	alveolar stop/alveolar trill release

Codes which are left free may be defined as necessary.

D - A segment synchronic property, transcribed by a subscribed diacritic, is coded in the most significant byte. The initial hexadecimal zero is a flag indicating that the first byte is to be interpreted as shown in Table 3:

y̰	0CB9	nasalized palatal approximant
z̪	0DA4	lateralized alveolar fricative
m̥	04C1	unvoiced bilabial nasal stop

Provision was made to code the lenis quality on a par with the fortis. However, the lenis quality is assumed to be the unmarked case and it is not associated with a graphic diacritic:

t	0114	lenis t
t̠	0214	fortis t

E - Consonants may be syllabic and bear tones. The syllabicity is coded by the least significant byte set to zero:

m̩	C100	syllabic bilabial nasal stop
s̩	9400	syllabic alveolar unvoiced fricative

Tones on consonants are coded as they are on vowels (see VOWELS, B); tone bearing consonants are assumed to be syllabic.

ḿ	C104	syllabic nasal stop/high tone
m̀	C102	syllabic nasal stop/low tone

F - The overwhelming majority of known consonants may be coded according to the preceeding conventions. However it may be crucial in some languages to handle difficult cases as accurately as possible. We shall resort to the following system: the most significant byte is used as a pointer to a specific filter corresponding to the primary consonant coded in the second byte. One has access, through this filter, to a complementary code, so that the resulting code is extended to 3 bytes; the flag set to detect this situation is the zero corresponding to the least significant bits of the first byte:

ndʳ Prenasalized stop/ trill release	10C4 filter C4/1 : 44F0 extended code : C444F0
ŋʗ̤ Murmured prenasalized click	10CB filter CB/1 : 0567 extended code : CB0567
ŋʗ Voiced prenasalized click	20CB filter CB/2 : 0467 extended code : CB0467

VOWELS

A - A short vowel - one mora - is coded on a short integer. A long vowel or a diphthong is coded as two morae. The most significant byte corresponds to segmental information. Vowels are plotted on an articulatory space defined by two axes: height (5 degrees) and tongue position in the oral cavity (front, central, back):

	Front	Central	Back
height	1	6	B
	2	7	C
	3	8	D
	4	9	E
	5	A	F

The most significant bits are interpreted as follows:

bit 0 - approximant-like vowel
1 - marked tongue root
2 - nasal
3 - round

The bit 0 is used to mark superclosed vowels (like reconstructed proto-bantu ɨ/ʉ) or, more generally, the non syllabic part of a diphthong:

əi̯	0A00	8100	diphthong with gliding i
iə̯	0100	8800	diphthong with gliding ə

The bit 1 is used to interpret marked tongue root position (emphatic vowels in the Berber-Arabic domain or the harmonic set of vowels characterized by Advanced Tongue Root in a number of subsaharan languages). Nasality and roundness may combine with this feature:

i	0100	(unrounded) i
u	1B00	(round) ʊ
i̟	2100	nasalized i
u̟	3B00	nasalized u
I̟	4200	ATR I

Basic symbols corresponding to the set of unrounded vowels and of rounded vowels are shown in Tables 4 and 5 respectively.

B - Suprasegmental information is coded in the second byte. Tonal languages use up to 5 levels of pitch, represented henceforth as accents. The code 06 is reserved for a downstepped High:

0101	Falling low	ì̀
0102	Level low	ì
0103	Mid	ī
0104	High	í
0105	Suprahigh	í̋
0106	Downstepped High	ꜜi

Contour tones are coded by reference to their source/target pitch:

0142	Falling high-low	î
0124	Rising low-high	ǐ

The bit 4 is set to 1 if the corresponding tone is floating:

014A	High + Floating low	í̀
012C	Low + Floating high	ì´

Double contours require two morae; we propose the convention that the first mora bear a level tone and the second a contour tone:

0104	0124	Falling-rising long i	íǐ
0102	0142	Rising-falling long i	ìî

C - In order to maximally compact suprasegmental information the bit 0 is reserved for stress:

0180	stressed i	ˈi

If the stressed vowel bears a tone, the code is modified accordingly:

0182	stressed i/low tone	ˈì
01C2	stressed i/falling tone	ˈî

The code A0 is assigned to pitch accent as required by some languages:

01A0	i associated with pitch accent	i̍

D - Hexadecimal codes 7 and F are left free in our system. Corresponding combinations will be used to account for marked voice quality:

unvoicing	0107	unvoiced i	i̥
	0147	unvoiced i high tone retained	í̥
creaky voice	012F	creaky i/low tone	ḭ̀
breathy voice	0172	breathy i/low tone	ì̤

Special cases may be treated with an extended code as proposed for consonants: a flag (hexadecimal F) indicates that one has to go through a filter table, access to which is given by the code of the vowel mora and a pointer:

01F2 : go to case 2 of the filter table corresponding to vowel i.

Rhotacized vowels, for instance, could be conveniently dealt with in this way.

CONCLUSION

It is indeed possible to rely on the International Phonetic Alphabet to propose a comprehensive and versatile computer oriented coding system. The fact that the code is phonetically motivated makes it particularly attractive for expert systems aiming at comparing data or reconstructing proto-languages.

Reference

[1] P. Ladefoged, "Preliminaries to Linguistic Phonetics", The Univ. of Chicago Press, 1971.

		bilabial 1	labiodental 2	dental 3	alveolar 4	labioalveolar 5	retroflex 6	postalveolar 7	prepalatal 8	palatal 9	labiopalatal A	velar B	labiovelar C	uvular D	pharyngeal E	glottal F
unvoiced consonants	1	p			t	pt	ʈ			c		k	kp	q		ʔ
aspirated	2	ph			th							kh				
ejectives	3	p'			t'							k'				
voiced consonants	4	b			d	bd	ɖ			ɟ		g	gb	ɢ		
implosives	5	ɓ			ɗ							ɠ				
clicks	6	ʘ	ʇ	ʖ			!	ʗ								
unvoiced affricates	7		pf		ts			tʃ					kf			
voiced affricates	8		bv		dz			dʒ					gv			
unvoiced fricatives	9	ɸ	f	θ	s		ʂ	ʃ	ɕ	ç		x		χ	ħ	h
voiced fricatives	A	β	v	ð	z		ʐ	ʒ	ʑ	j		ɣ		ʁ	ʕ	ɦ
approximants	B	ʋ			ɹ					y	ɥ		w			
nasals	C	m			n		ɳ			ɲ		ŋ		ɴ		
laterals	D				l		ɭ			ʎ		ɫ				
flaps /taps	E				ɾ											
trills	F	ʙ			r		ɽ							ʀ		

Table 1

Symbol	Example	Code	Phonetic interpretation
˺	t˺	1410	unreleased
ʰ	tsʰ	7420	aspirated release
ˀ	tsˀ	7430	glottal release
ʸ	kʸ	1B90	palatal release
ᶣ	kᶣ	1BA0	labiopalatal release
ⁿ	tⁿ	14B0	nasal release
ʷ	bʷ	41C0	labiovelar release
ˡ	tˡ	14D0	lateral release
ᶛ	tᶛ	14E0	pharyngeal release
ʳ	dʳ	44F0	trill release

Table 2

Symbol	Example	Code	Phonetic interpretation
	t	0114	lenis
̠	t̠	0214	fortis
̥	m̥	03C1	unvoicing
̬	s̬	0494	voicing
̤	b̤	0541	murmur
̫	s̫	0A94	rounding
˒	t˒	0B14	velarization
̰	w̰	0CBC	nasalization
̩	z̩	0DA4	lateralization
̣	ṭ	0E	pharyngalization
...	b...	0F41	laryngalization

Table 3 -

Unrounded vowels

i	ɨ	ɯ
ɪ	ᵻ	ɯ
e	ə	ɤ
ɛ	ɐ	ʌ
æ	a	ɑ

Table 4

Round vowels

y	ʉ	u
ʏ	ʉ	ʊ
ø	ɵ	o
œ	ɞ	ɔ
ɶ	ɒ	ɒ

Table 5