PHONETIC CODING FOR DATA BASES AND EXPERT SYSTEMS

GILBERT PUECH et PIERRE BANCEL
Centre de Recherches Linguistiques et Séméiologiques
Université Lumière-Lyon 2
69500 Bron, France

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 [l] or [l] is exclusive of [e] - or hardly satisfactory to account for some sounds - such as glus and trills. On the other hand an IPA based classification presents several advantages: it is independent 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 codes. It should also be used to classify phonetically recorded words in the data base 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 algorithms 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 a hexadecimal figure. Consonants and vowels are coded independently 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 Sante - in which words are built after a strict syllabic pattern, the data base may determine the fields composing the record as corresponding either to a consonant or to a vowel; in languages where no such syllabic regularity prevails, the first field of the record (long int) will in the first byte determine the nature of segments included in the record and, in the three following bytes, select the V/C choice (bit 0 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 minimal versatility to this coding system which was designed both for narrow and broad transcriptions. Coding of morphone boundaries for morphophonemic representations was not examined but could be accommodated.

Consonants

- 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 9 most significant bits correspond to the lines (sense of articulation) and the n remaining bits to the columns (place of articulation):}

<table>
<thead>
<tr>
<th>Phonetic symbol</th>
<th>Code</th>
<th>Phonetic interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>0001</td>
<td>bilabial voiced stop</td>
</tr>
<tr>
<td>m</td>
<td>0011</td>
<td>bilabial nasal stop</td>
</tr>
<tr>
<td>k</td>
<td>0101</td>
<td>labiodental unaspirated stop</td>
</tr>
<tr>
<td>kp</td>
<td>0102</td>
<td>labiodental unaspirated stop</td>
</tr>
</tbody>
</table>

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

Vowels

- A short vowel - one mora - is coded on a short integer. A vowel or a diphthong is coded as two morae. The most significant byte is interpreted as a vowel and the second as 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 minimal versatility to this coding system which was designed both for narrow and broad transcriptions. Coding of morphone boundaries for morphophonemic representations was not examined but could be accommodated.

Consonants

- 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 9 most significant bits correspond to the lines (sense of articulation) and the n remaining bits to the columns (place of articulation):
the non syllabic part of a diphone:

adjacent

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

1. 0100 (unrounded) 1
2. 1000 (rounded) 1
3. 0101 (nasalized) 1
4. 0102 (nasalized) u
5. 0103 (ATR) 1

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

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 hereafter as accents. The responding to VOWELs laterals D A

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

0102 stressed i/low tone 1
0107 stressed i/low falling tone 1

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

0100 1 associated with pitch accent 1

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

unvoiced 0107 unvoiced 1
unvoiced 0147 unvoiced 1
creaky voice 012F unvoiced 1
breathy voice 012B breathy 1/low tone 1

Special cases may be treated with an extended set 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 in the first mora and a pointer:

012F : go to case 2 of the filter table corresponding to vowel 1.

Rhotastic 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


Symbol Example Code Phonetic interpretation
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis
- " t 1410 aspirated release - " t 0114 lenis