ARCHISEGMENT-BASED LETTER-TO-PHONE CONVERSION FOR CONCATENATIVE SPEECH SYNTHESIS IN PORTUGUESE

Eleonora Cavalcante Albano and Agnaldo Antonio Moreira
LAFAPE-IEL-UNICAMP, Campinas, SP, Brazil
albano@ccsun.unicamp.br

ABSTRACT

A letter-to-phone conversion scheme is proposed for Portuguese which excludes representation of allophonic detail. Phonetically unstable segments are treated as archisegments, their articulatory weakness being analyzed in terms of feature underspecification. Besides solving classical problems of allophony and allomorphy, this analysis provides an efficient principle for building a unit inventory for concatenative speech synthesis.

1. PHONOLOGY AND LETTER-TO-PHONE CONVERSION

Concatenative speech synthesis depends crucially on the adequacy of the phonological analysis underlying its unit list. In the same way as intelligibility requires concatenative units to be based on a consistent minimal set of allophones, quality requires enriching the allophone inventory. Allophony, however, is an old difficulty in phonological theory. Since the 50's the literature has been discussing whether there is a clear line separating phonemes from allophones [1], or, in today's terms, phonological from phonetic segments. Contemporary laboratory research has, in addition, uncovered many cases of allophony where the variants, rather than falling into discrete categories, vary continuously along some phonetic dimension, making it difficult to talk about segments at the phonetic level [2].

From the point of view of text-to-speech conversion these theoretical issues lead to persistent practical difficulties concerning the size of the phone inventory and the number of steps in letter-to-phone conversion. In concatenative systems, such questions may also affect the determination of unit size, since phonetically unstable allophones tend to coarticulate with more than one adjacent phone, thus requiring units larger than the diphone in order to be successfully represented.

In our laboratory, we have been dealing with these problems as part of an effort to develop a high quality concatenative speech synthesis system for Brazilian Portuguese. An earlier attempt within our own research group [3] has yielded a low cost, low quality TD-PSOLA system. Since our aim is academic, not commercial, we are now laying emphasis on the kinds of improvement that may be achieved through the solution of theoretical problems in the analysis of language and speech.

2. THE PORTUGUESE PROBLEM

Portuguese orthography is quasi-phonemic and thus very easy to convert to phones except in a few cases where allophony is extensive. Interestingly, such exceptions occur precisely where complexity at both the phonetic and the morphological level leads to ambiguity in phonological analysis.

Take, for example, the representation of rhyme nasality, which motivates the controversy whether Portuguese has distinctive nasal vowels [4,5]. Such nasality, which tends to be rather heavy phonetically, is orthographically represented as 'm' or 'n' before a word internal consonant (e.g., *samba*, *santa*, *sanca*) and as a tilde diacritic on the vowel before another vowel or word finally (e.g., *são*, *sã*). This cannot be said to be a reasonable phonemic representation because /m/ and /n/ do not contrast in rhymes, a fact that orthography acknowledges by making the choice of one or the other letter dependent on the following consonant ('m' before 'p,b' and 'n' elsewhere). Nor can it be said to be a reasonable allophonic representation because the phonetic realization of nasal rhymes shows no correspondence with the diacritic/digraph distinction.

Acoustic phonetic studies [6,7] have in fact shown that such rhymes may surface as nasal vowels or as nasalized vowels followed by nasal murmurs, regardless of the orthographic representation. In our laboratory, we have, moreover, observed that this murmur has a variable length, ranging from negligible to sizable. Such variation evokes the continuous segment reduction patterns discussed by Sproat and Fujimura [8] and Kohler [9], among others, being thus likely to be conditioned by higher level factors, such as speaker, speech rate, accent contour, and style.

Facts such as this argue against phonetic detail as the proper level for letter-to-phone conversion. Any transcription of Portuguese aiming at distinguishing between nasal rhymes pronounced with and without a nasal murmur would probably have to deal with the interaction between segments and prosody, a question which is complex from a theoretical point of view and costly from a computational point of view.

Other cases where allophones cannot be stated from simple context-sensitive rules exist in Portuguese [10] and are also attested in other languages [11]. To deal with them, letter-to-phone conversion has to aim at a relatively abstract level of analysis. The ensuing question recapitulates the major concern of the phonology of the 70's [12]: how abstract should the proper level of representation be?

3. AN ARCHISEGMENTAL SOLUTION

The answer we propose for Portuguese is just abstract enough to solve problems at both the phonetic and the morphological level without giving up the convenience of generating a linear string of

symbols which can be easily matched with orthography. It consists of treating all the segments that show extensive allophonic and allomorphic variation as underlying archisegments (i.e., underspecified segments) and exploring this distinction all the way down to the unit segmentation algorithm. To implement this view, we perform letter-to-phone conversion in two steps. One works on words and has an output that may be interpreted as a shorthand for a lexical phonological representation. The other level where coarticulation becomes really crucial, i.e., unit segmentation (see section 5).

Table 1 summarizes our analysis of the segmental phonology of Portuguese and its orthographic correspondences. Conversion rules are quite straightforward. An error rate of 4% is handled in an exception dictionary.

Phonology	Orthography	Phonology	Orthography	Phonology	Orthography	Phonology	Orthography
p	p	S	s, z, x	NI	nh	a	a
b	b	n	n	LI	lh	oh [ɔ]	o
f	f	N	m, n, ã, õ	k	c, qu	o	o
v	v	l	l	g	g, gu	u	u
m	m	L	l			I	e, i
t	t	r	r, rr			E	e
d	d	R	r	i	i	A	a
s	s, ss, c, ç	sh [ʃ]	ch, x	e	e	O	o
z	s, z, x	zh [ʒ]	j, g	eh [ɛ]	e	U	o, u

Table 1 - Archisegmental notation for Portuguese letter-to-phone conversion.

works on sentences and has an output that may be interpreted as a shorthand for a postlexical phonological representation.

Portuguese has several segments which not only have a widely dispersed range of allophones but also participate in allomorphy associated with inflection and derivation. Our approach to this problem is to assume that both allophony and allomorphy implement segment strength distinctions which hinge on feature specification. Allomorphy is a lexical mechanism which adds or deletes features. Allophony is a language specific coarticulation mechanism which exerts itself in the inverse proportion of the number of specified features. In this perspective, allophony is not represented in letter-to-phone conversion except as the capital letter (archisegment) notation assigned to highly variable segments.

For example, our representation for nasal rhymes is uniformly /VN/ (i.e., vowel + nasal archisegment). Thus, the so-called nasal diphthongs *ão, ãe, õe* are represented as /aNU, aNI, oNI/. This analysis is consistent with the fact that such sequences occur only at ends of base words, their /N/ being replaced by /n/ in derived words (see section 4). Whether such rhymes will be more vocalic or more consonantal (i.e., will have a nasal murmur or not) depends on phonetic implementation and does not matter until the

The main purpose of this notation is to represent lexical and postlexical distinctions while leaving the phonetic realization of highly variable segments relatively open. This is achieved by distinguishing between two degrees of segment strength, which are translated into two degrees of feature specification. Segments having relatively stable phonetic realizations are treated as strong and hence fully specified. They are: primary onset consonants (p, b, f, v, m, t, d, s, z, n, l, r, sh, zh, k, g), vowels preceding lexical stress (i, e, a, o, u), and vowels bearing lexical stress (i, e, eh, a, oh, o, u). Segments having unstable phonetic realizations are treated as weak and hence underspecified (archisegments). They are: rhyme consonants (N, L, R, S), secondary consonants in onset clusters (L, R), sonorants restricted to intervocalic position (R, LI, NI), semivowels (I, U), and vowels following lexical stress (I, E, A, O, U). Note that palatal sonorants are analyzed as clusters (NI, LI).

4. EVIDENCE

Let us now briefly review the linguistic evidence for this analysis, which is more fully developed in [10].

Firstly, it dispenses with stress marks by treating stressability as a function of feature specification or strength. Instead of distinguishing between stressed and unstressed vowels, it

distinguishes between vowels that may be accented postlexically - namely, those going from the beginning of the word to the so-called "lexical stress" (e.g. *casar* /kazaR/) - and vowels that may not - namely, those following the latter: the so-called "post-stressed" vowels (e.g. *casa* /kazA/). Potential accent bearers are strong and fully specified. Non-accent-bearers are weak and underspecified.

Secondly, it accounts for several cases of allomorphy in inflection and derivation. For example, *pão* /paNU/, 'bread, singular', alternates not only with *pães* /paNIS/, 'bread, plural', but also with derived verb *panificar* /panifikaR/, 'to make bread'. The singular/plural pair is derived from the root /paN/ through addition of the empty inflectional vowel /V/, which turns into the default labial vowel /U/ in final position and into the default high vowel /I/ before /S/. *Panificar*, is, in turn, derived from the root /paN/ through addition of the coronal feature, which turns /N/ into /n/. This strengthening of /N/ is triggered by the presence of the strong, fully specified vowel /i/ at the beginning of the derivational suffix /if+ik+aR/. Similarly, *sol* /sɔL/, 'sun, singular', alternates with *sóis* /sɔIS/ 'sun, plural' and derived adjective *solar* /solaR/, 'solar'. In the same vein, this is explained by positing an underlyingly underspecified lateral at the end of the root /soL/, which may weaken and fall before a weak vowel (here /I/) and strengthen and turn into a coronal lateral before a strong vowel (here /a/).

Thirdly, it expresses important differences in syllabification without any actual syllable parsing. For example, disyllabic *baia* /baIA/, 'stall', is distinct from trisyllabic *baía* /baiA/, 'bay', because weak high vowels are realized as glides in intervocalic position while strong high vowels form heterosyllabic clusters. Analogously, the closed initial syllable of *abrogar* /abrogaR/, 'to abrogate', is distinct from the open initial syllable of *abrolhos* /abRohLIOS/ , 'reefs', because the weak rhotic /R/ obligatorily forms an onset cluster with a preceding obstruent while the strong rhotic /r/ does not.

Fourthly, it draws a sharp line between postlexical and phonetic processes by assuming that the former do not introduce any segmental symbols beyond those already in the lexicon. For example, in *vem aqui* /veNIaki/, the postlexical link between /veN/ and /aki/ is established by /I/, a vowel which already exists in the lexical inventory. By contrast, processes which apparently create new segment classes, such as the pronunciation of *ti* /ti/, 'you, oblique case', as [tʃi], are taken to be phonetic (i.e., continuous coarticulation phenomena), and thus are not represented in symbolic terms.

This should suffice to give an overview of the linguistic framework behind Table 1. The main point here is that the distinction between strong and weak segments, which is formally expressed through feature specification, affords a great deal of symbolic economy at the lexical and postlexical levels, besides signaling non-symbolic differences which are realized as different degrees of coarticulation at the phonetic level.

5. CONSEQUENCES FOR CONCATENATIVE SYNTHESIS

In addition to simplifying letter-to-phone conversion, this analysis implies an extremely simple principle for assembling a unit inventory for concatenative speech synthesis, namely: archisegments - or weak segments -, which tend to be highly coarticulated with their context, should not be separated from adjacent fully specified segments.

Our first attempt to apply this principle yielded an inventory of over 10,000 units, which is beyond the current capability of our laboratory. This led us to look into a weaker version of the same principle, which says that archisegments should not be separated from adjacent fully specified segments *except where constituting syllable nuclei*. This modification, which eliminates 7,616 polyphones centered around the weak vowel nuclei /I,E,A,O,U/, keeps the inventory down to 3,537 units.

A step-by-step example will help illustrate how the analysis works at the levels of letter-to-phone conversion and unit segmentation. Consider the sentence:

(1) *Este é um sintetizador brasileiro.* 'This is a Brazilian synthesizer.'

The first pass of letter-to-phone conversion (lexical level) yields:

(2) //eStE eh uN siNtetizadoR bRazileIRO//

The second pass of letter-to-phone conversion (postlexical level) applies boundary related rules (e.g., raising of final /E, O/) and introduces more prominence contrasts by cliticizing grammatical words (e.g., /uN/ becomes /UN/) and weakening every other vowel to the left of the last strong vowel of polysyllables ending in weak vowel (e.g., /zi/ becomes /zI/):

(3) //eStIehUNsiNtetizadoRbRazIleIRU//

This string is finally parsed by the unit segmentation algorithm as follows:

(4) //e- eSt- tI- Ieh- ehU- UNs- si- iNt- te- et- ti- iz- za- ad- do- oRb- bRa- az- zI-Il- le- eIRU- U//

This parsing follows the weak version of the above proposed principle, which has two important consequences for the phonetic content of the units. Firstly, all vowels in syllable onset units (i.e., CV or CCV) are treated as oral, which requires that the domain of nasality in Portuguese be unmistakably the rhyme. Secondly, as a result of the weakened version of the principle, archisegments in syllable nuclei, differently from those in syllable margins, are split in the middle, which requires that formant trajectories in the vicinity of the vowel target be either stable or very short. Note that, while stability would contradict the non-stationariness attributed to archisegments in the present framework, shortness would not.

Fortunately for the economy of our system, both predictions are born out by phonetic fact. Acoustic phonetic analyses have shown that (a) nasal vowels are much more nasalized at the end than at the beginning [13]; and (b) weak vowels are best described as an intersection of transitions from adjacent segments, without any significant straight line in the middle [14].

In addition, exploratory concatenation experiments based on the ensuing segmentation algorithm have yielded very good auditory results. In spite of the absence of prosodic processing, concatenated sentences exhibit not only the right lexical stress pattern but also a natural, agreeable phrasal rhythm.

Encouraged by these results, we have set out to record and segment the 3,537 unit inventory, and hope to be able to report on a TD-PSOLA system based on the above analysis by next year.

6. REFERENCES

1. Halle, M. *The sound pattern of Russian*, Mouton, The Hague, 1959.

2. Pierrehumbert, J. and D. Talkin "Lenition of /h/ and glottal stop," Docherty, G. and R. Ladd (eds.) *Papers on Laboratory Phonology II: gesture, segment, prosody*. Cambridge University Press, Cambridge, 1992.

3. Egashira, F. and F. Violaro "Conversor texto-fala para a língua portuguesa," presented at *13º Simpósio Brasileiro de Telecomunicações*, Águas de Lindóia, São Paulo, 3-6 September 1995.

4. Viana, A. G. "Essai de phonétique et de phonologie de la langue portugaise d' après le dialecte actuel de Lisbonne," *Estudos de fonética portuguesa*. Imprensa Nacional, Lisboa, 1973[1883].

5. Câmara Jr., J. M. *The Portuguese language*, University of Chicago Press, Chicago, 1972.

6. Almeida, A. "The nasal vowels: phonetics and phonemics," Schmidt-Radefelt, J. (ed.) *Readings in Portuguese linguistics*. North Holland, Amsterdam, 1976, 349-287.

7. Barbosa, J. M. "Les voyelles nasales du portugais," In: *Proceedings of the 4th International Congress of Phonetic Sciences*, Mouton, The Hague, 1962, 691-709.

8. Sproat, R. & O. Fujimura "Allophonic variation in English /l/ and its implications for phonetic implementation", *Journal of Phonetics* 21: 291-311, 1993.

9. Kiparsky, P. "Abstractness, opacity and global rules," Fujimura, O. (ed.) *Three dimensions of linguistic theory*, TEC Company, Tokyo, 1973, 57-86.

10. Albano, E. C. "Demarcative feature specification in phonology and phonetics: the case of portuguese allophony/allomorphy", presented at the *Fifth Conference on Laboratory Phonology*, Evanston, Illinois, 6-8 July 1996.

11. Choi, J. D. "An acoustic-phonetic underspecification account of Marshallese vowel allophony," *Journal of Phonetics*, 23: 323-327, 1995.

12. Kohler, K. J. "Articulatory reduction in different speaking styles," *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, 1995, vol. 2: 12-19.

13. Sousa, E. G. "Para a caracterização fonético-acústica da nasalidade no português do Brasil," master's thesis. LAFAPE, IEL-UNICAMP, Campinas, 1994.

14. Aquino, P. A "Redução de vogais em ambientes lexicais em português brasileiro," master's thesis. LAFAPE, IEL-UNICAMP, Campinas, forthcoming.

Acknowledgements: Research supported by **FAPESP**, grant no. 93-0565-2, and **CNPq**, grant no. 52.335/94-6.