LINGUISTIC CRITERIA FOR BUILDING AND RECORDING UNITS FOR CONCATENATIVE SPEECH SYNTHESIS IN BRAZILIAN PORTUGUESE

Eleonora Cavalcante Albano and Patrícia Aparecida Aquino
Laboratório de Fonética Acústica & Psicolingüística Experimental (LAFAPE)
CP 6045, IEL-UNICAMP,
Campinas, SP, 13081-970, Brazil.
Tel: +55 19 2398596, FAX:+55 19 2391501, E-mail: albano@iel.unicamp.br

ABSTRACT

A unit inventory for concatenative speech synthesis in Brazilian Portuguese was built on the basis of an analysis of segment-prosody interactions. Segments are viewed as full or reduced depending on stress, syllable structure and phonological boundaries. Demisyllabic units preserve the integrity of segments reduced due to syllable structure. Intersyllabic units preserve the integrity of segments reduced due to stress and boundaries. Integrity of vowel clusters is also preserved, but nasal vowels and diphthongs are successfully concatenated to oral onsets. The resulting units were recorded in carrier words and sentences designed on phonotactic and grammatical grounds. Good quality concatenation is achieved even before the addition of prosodic treatment.

1. INTRODUCTION

This paper describes the phonological and grammatical criteria used for creating a unit inventory for concatenative speech synthesis in Brazilian Portuguese (henceforth BP) and choosing the carrier words and sentences to embed them for recording and excising.

On the one hand, BP challenges speech synthesis in that it has many vowel clusters and various segment weakening processes which have a complex input (generally a combination of prosodic and stylistic factors) and an extremely variable output, ranging over continuous rather than discrete phonetic dimensions (for example, reduced vowels shift continuosuly over the vowel space). On the other hand, BP encourages intonation synthesis, inasmuch as it allows for relatively few intonation phrases per sentence and implements pitch accents with relatively small f_0 excursions. Its tone of voice is rather flat as compared, for instance, to that of English or Spanish. Consequently, a concatenated declarative with no f_0 or duration adjustment can sound quite right if word stress and stress-related allophony are correctly placed. A unit inventory based on analysis of segment-prosody interactions is thus a major step for achieving quality speech synthesis in BP.

In our view, most of the segmental variability of BP is due to syllable structure, stress, and word or phrase boundaries. To capture such effects, we have built an inventory combining demisyllabic units [1], which preserve most syllable internal coarticulation, with intersyllabic units, which add coarticulatory effects due to syllable and word boundaries. The total number of units, which range from diphones to tetraphones, is 2,193. Speech concatenated from this inventory sounds quite natural even in the absence of a prosodic module, which is still under construction [2].

2. THE PHONOLOGICAL ANALYSIS

Allophones are hard to specify in BP because their actual phonetic realization depends on variable factors such as dialect, style and strength of prosodic boundary. A concrete segment inventory, corresponding roughly to a phonetic transcription, would imply many arbitrary choices of prosodic and stylistic norms, besides depending on dialect or even on speaker. An abstract segment inventory, corresponding roughly to a phonemic transcription, would in turn miss crucial contextual variability.

In order to solve this problem, we have elaborated an abstract yet prosodically differentiated segment inventory. Segments which in a traditional analysis would correspond to a single phoneme with different allophones appear in two versions: the full one, occurring in prosodically strong environments, and the reduced one, occurring in prosodically weak environments. Reduced segments are weak and phonetically more variable than full ones. The reducing environments are: edges of syllable constituents and position relative to stress. Reduced consonants and vowels occur at the edges of onsets, nuclei and rhymes. Reduced vowels also occur in the nuclei of syllables following word stress (two at most).

This analysis considerably simplifies letter-to-phone conversion, as it creates an archisegmental (i.e., only partially specified) phonetic notation which directly expresses segment-prosody interactions, making it uncessary to mark stress and syllable boundaries [3]. The ensuing 34 segments can be displayed and grouped as shown below (reduction is marked by upper case)

It should be noted that these are more properly termed *autosegments* [4], since there are cases in which two share a single syllable constituent slot (e.g., in (h) and (i) below, N shares the weak nucleus slot with I/U and I, respectively, to form the nasal diphthongs aNI, aNU, oNI and eNI).

Full Consonants

	labials	coronals	pal./vel.
stops	p	t	k
	b	d	g
fricatives	f	s	sh
	v	z	zh
nasals	m	n	nh
laterals		l	lh
rhotics		r	

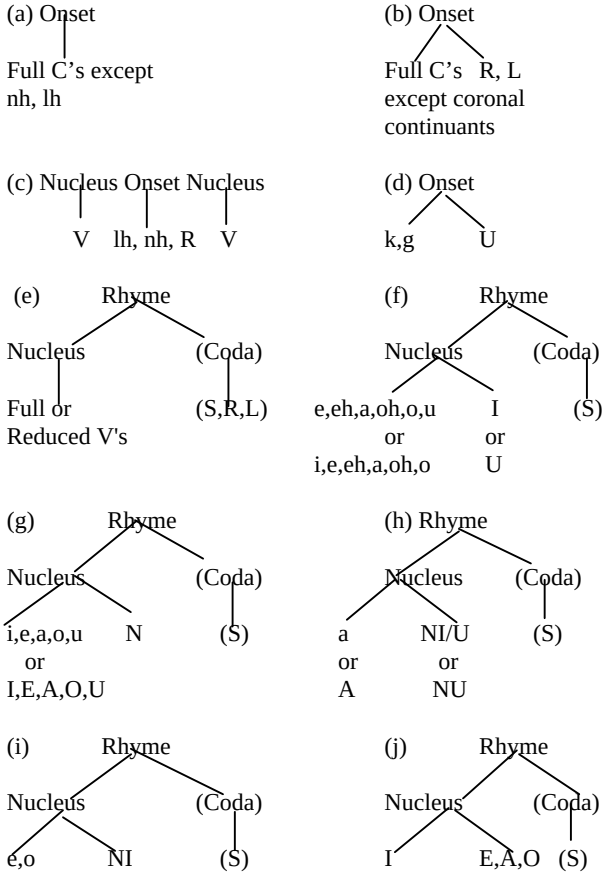
Full vowels

i u
e o
eh oh
a

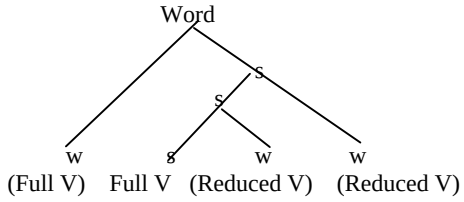
Reduced Consonants: S

Reduced Vowels: N i u; L e o; R a

The syllable level constraints on the occurrence of reduced vowels and consonants can be expressed by the following demisyllable trees:



The word level constraint on the occurrence of reduced vowels can be expressed by the following metrical tree:



An obvious consequence of the above analysis is that segmenting the speech signal in reducing environments should lead to poor concatenation. Thus, the ideal inventory for concatenative speech synthesis in BP would allow the waveform to be cut only at stationary points of full segments. But this would lead to a proliferation of long units in order to accommodate all the contexts of occurrence of reduced segments. Take, for example, reduced complex nuclei followed by coda, such as in (h) or (j) above. A constraint against segmenting reduced segments would yield units as long as six autosegments (e.g., *orgãos maus*, 'bad organs', [ohRgANUSmaUs], segmented into ohRg-gANUSm-ma-aUS). Estimation from such complex cases shows that the inventory size under this perspective would be around 20,000 units.

Since managing such a large *corpus* is beyond the means of our laboratory - a small teaching and research facility in a public university -, we have resorted to phonetic analysis in search of criteria to cut down on the inventory without loss of quality.

3. THE PHONETIC ANALYSIS

From the above discussion the cases that strike as potentially problematic concern reduced vowel nuclei, complex rhymes (including nasalized ones) and intervocalic [R].

The figures for reduced vowel nuclei are disturbing: embedding all possible reduced vowel nuclei in all possible preceding and following onsets would alone yield a 8,000 unit set, which would double with the addition of the cases where such nuclei are followed by [S], itself a reduced segment.

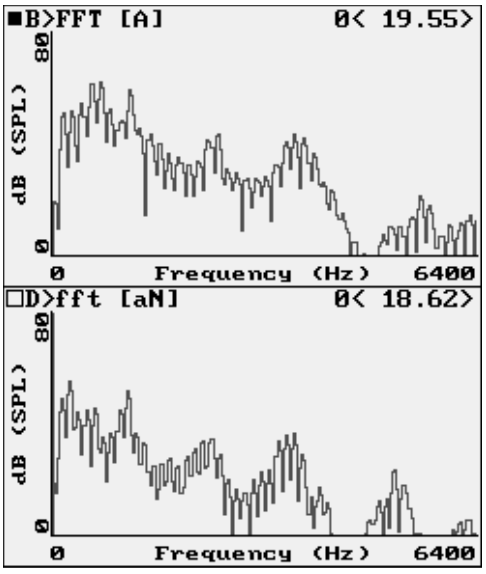
The question clearly arises whether reduced nuclei lend themselves to concatenation, in spite of the fact that their F-pattern is seldom stationary.

Since we had anyway planned to have intersyllabic units accounting for coarticulation of nuclei with the following onsets, we tried to perform concatenation at the very edge of demisyllables, i. e., at the point where the transition from the preceding consonant ends. The results were highly satisfactory: spectral discontinuities are much less perceptible early in nuclei than halfway. Thus, concatenation following the demisyllabic principle even when the second "half" of the demisyllable is actually intersyllabic led to a drastic reduction of the inventory built around stressless vowels: the figures came down to 746.

Demisyllabic concatenation was also helpful in solving a substantial part of the complex nuclei problem. As in other languages, BP diphthongs cannot be broken on account of rapid spectral change. So, oral diphthongs generally require triphones: the diphthong itself plus a following segment. But what about nasal vowels and diphthongs? Do they require special onsets or can nasal rhymes be concatenated to oral onsets? To take an example, should *bom* [boN] 'good' be segmented as [bo-oN] or [boN-oN]?

Descriptive work conducted at our laboratory had already offered support to the first hypothesis by showing that the acoustic indices of nasality show up gradually in the rhyme, being much stronger at the end than at the beginning [5]. Informal experiments with our current informant (the owner of the synthesizer's voice) confirmed this view: concatenating onsets excised from oral contexts with nasal rhymes gave quite good results except in the case of [aN, AN, aNU, aNI, ANU].

We did not, however, have to give up the economy afforded by demisyllabic concatenation. As the problem arises from a discrepancy in vowel quality due to the fact that nasalization raises the low vowel making it sound like an a-colored schwa, we simply tried to substitute the reduced vowel for the full one in those onsets to be concatenated to [aN] and the like. This worked very well because [A] is actually very similar to [aN] in oral formant pattern, as shown by the spectra below:



So, in order to generalize the rule concatenating oral onsets with nasal rhymes, the only necessary adjustment was to add a rule to the unit segmentation algorithm turning the left [a] into [A] as [aN] is broken into two demisyllabic units. For example, *mão* [maNU] 'hand' is rewritten as [mA - aNU].

Demisyllabic concatenation provided good solutions to these cases but not to that of intervocalic [R]. This is because this segment is generally realized as a very short tap (mean duration around 15 ms), with a closure highly coarticulated with the preceding or the following vowel, depending on stress [6]. Concatenation in the tap closure is often complicated by pitch marking errors due to disturbances of voicing during the closure. In addition, the resulting spectral discontinuities give a reverberating quality to the signal.

Fortunately, intervocalic [R] is fairly restricted: it combines only with oral vowels and diphthongs, so that the number of intersyllabic units needed to embed it into triphones is under 200.

Another case in which phonetic analysis helped sharpen the inventory was that of coda [L]. While some BP dialects have a velar lateral in coda, others have a velar glide in its place, and still others alternate stylistically between the two. A study of a sentence *corpus* read by our speaker placed him securely in the second type of dialect. We were thus able to save 284 units by adding a rule turning coda [L] into [U] to the letter-to-phone converter.

Finally, further analysis of the same *corpus* showed that the quality of pre-stressed vowels is closer to that of stressed vowels than to that of post-stressed vowels, which supports treating them as full, as in the metrical tree of section 2. Though some reduction does take place in pre-stressed position as well, the only effect of ignoring it is to make the speech sound slightly hyperarticulated, which may be desirable for synthesis purposes. There remains, however, the possibilility of resorting to the reduced vowel set in polysyllabic environments, where pre-stressed position may be more susceptible to reduction, requiring an alternation between full and reduced vowels. Implementing this feature is just a matter of adding a few rules to the letter-to-phone converter.

4. THE UNIT INVENTORY

The construction of the inventory is based on two principles: the distinction between full and reduced vowels and the integrity of vowel clusters. Thus, three unit sets were designed: one around full nuclei, another around reduced nuclei, and another around permissible vowel clusters.

The full and the reduced set are very similar in structure. The first includes ten types of units: onset-nucleus (e.g, pa, pRa), nucleus-onset (e.g., ap, aNU#p), nucleus-coda (e.g., aS), coda-onset (e.g., Sp), nucleus-onset-nucleus (e.g., aRA), nucleus-coda-nucleus (e.g., aS#a), nucleus-silence (e.g., a//), nucleus-coda-silence (e.g., aS//), silence-onset (e.g., //s) and silence-nucleus (e.g., //a). The second includes all but the last two, since reduced vowels do not occur in utterance initial position.

Originally, we had planned to use nucleus-coda-onset units (e.g., aSp) instead of combining nucleus-coda (i.e., aS) and coda-onset (i.e., Sp) units as shown above. But since breaking heterosyllabic consonant clusters saves over 800 units, we made the relevant tests and concluded that the economy was worth the quality loss, which was indeed very small in this case.

Nucleus-onset-nucleus units were set up to deal exclusively with intervocalic [R], as explained above.

Nucleus-coda-onset units were introduced to handle final [S] at junctures with words beginning with vowels. On the surface, this [S] is voiced and sounds like [z], but, in most cases, the voicing is only partial and this fact plays an important role in signaling the word boundary. Since, in our opinion, adequate rendition of boundary allophones is a major key to naturalness, quality outweighs economy in this case.

Vowel clusters require two kinds of units: simple nucleus-simple nucleus (e.g., iA, ia) and complex nucleus-simple nucleus (e.g., oIA, aNU#a). These are in turn divided into restricted sets that occur only word internally (e.g., iA) and less restricted combinations that may occur at boundaries (e.g., Ai). Clusters occupying more than three syllable constituent slots (e.g., *são aias*, '(they) are servants', [saNU#aIAS], are handled through demisyllabic concatenation: [sA-aNU#a-aIA-AS].

Excising nucleus-nucleus units is not a trivial matter. As vowel clusters differ from diphthongs only in the slower rate of change of formant trajectories, it is sometimes tricky to determine the point where such trajectories stabilize so as to make a cut compatible with demisyllabic concatenation (i.e., at the end of the "empty onset"). Apparently, this can only be handled successfully in manual terms, i.e., by combining auditory monitoring with inspection of spectrograms and LPC formant tracks.

With these three sets, which total 2,193 units, any BP sentence, including those containing sequences of vowel clusters can be concatenated with fairly good quality. The following example shows that the longer units, which would be useful in any variety of Portuguese, are particularly so in BP, given the influence of native Brazilian languages: *O rio Araguaia separa Goiás do Pará*, 'the Araguaia river separates Goiás from Pará': [//o-o#r-ri-iO-O#a-aRa-ag-gUa-aIA-A#s-se-ep-pa-aRA-A#g-go-oIa-aS-S#d-do-o#p-pa-aRa-a//].

5. THE *CORPUS*

Many of the units discussed above do not occur within words and were set up explicitly to deal with word boundaries. The recording contexts were accordingly designed to reflect the distinction between units with and without obligatory boundaries.

Units that can occur word internally were recorded in nonsense words of the form *pa+unit* inserted in the carrier sentence *Digo _ baixinho* (I say _ softly).

Units that can occur in sentence initial and sentence final position were recorded, respectively, in nonsense words of the forms *unit+pa* and *pa+unit*, embedded in the carrier sentences: _ *digo baixinho* ('_ I say softly') and *Baixinho digo* _ ('Softly I say _').

To avoid violation of phonotactic constraints, nonsense words were not used with units that contain obligatory boundaries. These were instead embedded in word sequences of the form N + Adj , which were inserted in the carrier sentence *Este é um* _ ('This is a _'). The N + Adj construction is appropriate for this purpose because the syntactic boundary between the noun and the adjective is just strong enough to trigger sandhi phenomena but not silent pauses.

The choice of the noun-adjective pairs took into account stress pattern, word length, and, whenever possible, segmental context and word frequency. The result was a very natural pronunciation of the boundary units, which would have been impossible without real words.

6. CONCLUSION

If, on the one hand, the current experiment shows the importance of linguistic analysis in designing speech synthesis systems, on the other it shows the power of speech technologies in testing linguistic analyses.

The success of our inventory supports the hypothesis that most of BP allophony derives from stress, syllable constituency and phonological boundaries. It also supports the claim that a prosodically differentiated archisegmental notation is superior to an allophonic one in representing phonetic variability [3].

In addition, our work suggests that concatenative synthesis is the best available means for capturing low level phonetic phenomena that cannot as yet be modeled. The allophones represented in our inventory, which vary widely over several phonetic continua, would certainly sound much more schematic if modeled by a synthesis-by-rule system.

ACKNOWLEDGEMENTS

This research was funded by CNPq grant number 523555/94-6 and by FAPESP grant number 93/0565-2.

REFERENCES

[1] Fujimura, O. and J. Lovins "Syllables as concatenative phonetic units", In A. Bell & J. Hooper (eds.) *Syllables and segments*. Amsterdam: North Holland, 107-120, 1978.

[2] Barbosa, P. A. "A model of segment (and pause) duration for Brazilian Portuguese text-to-speech synthesis", *Proceedings* EUROSPEECH'97, 1997.

[3] Albano, E. and A. Moreira "Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese", *Proceedings of ICSLP 96*, vol 3, pp. 1708-1711, 1996 .

[4] Goldsmith, J. *Autosegmental and metrical phonology*. London: Blackwell, 1990.

[5] Sousa, E. M. G. "Para a caracterização fonético-acústica da nasalidade no português do Brasil", unpublished master's thesis, LAFAPE-IEL-UNICAMP, 1994.

[6] Silva, A. H. P. "Para a descrição fonético-acústica das líquidas no português brasileiro: dados de um informante paulistano", unpublished master's thesis, LAFAPE-IEL-UNICAMP, 1996.