Length and identification of Brazilian Portuguese segments in continuous speech signal

Leticia Rebollo†, Filipe Barbosa‡ , Rachel Costa†, Rosângela Lannes†, Josée Nicolau†, Carlota Rosa† and Fernando Resende ‡

† Faculdade de Letras, Universidade Federal do Rio de Janeiro, Brazil

‡ Escola Politécnica, Universidade Federal do Rio de Janeiro, Brazil

E-mail: leticiarcouto@yahoo.fr , filipe@lps.ufrj.br , rachelidyane@ig.com.br , rosangelannes@bol.com.br , joseenicolau@ig.com.br , carlota@centroin.com.br, gil@lps.ufrj.br

ABSTRACT

This paper identifies the pertinent acoustic segments to be applied in a Brazilian Portuguese text-to-speech synthesis system. In order to obtain a normalized length measurement, the vocalic segments are analyzed with respect to the lexical and phrasal stresses, besides its syllable positions in the initial and at the end of the utterances. Results show that these features are relevant factors to establish the acoustic units' duration mean.

1. INTRODUCTION

The aim of this work is to describe the relevant acoustic segments for a text-to-speech (TTS) synthesis system based on Portuguese spoken in Rio de Janeiro.

To synthesize speech, it is necessary to have a duration modeling of the acoustic units. This paper provides data that plays an important role in this research. In order to obtain the duration means for a posterior speech processing, length segments are analyzed in relation to the speech rate, the stress status and the utterance position.

This article is organized as follows: in Section 2 we present the speech database and the table of 76 vocal segments that are used in the segmentation; in Section 3 we describe the vocal segment means in relation to the primary and secondary stresses and its position in beginning or at the end of utterances; in Section 4 we introduce a discussion about duration modeling for the Brazilian Portuguese (BP) and Section 5 brings our conclusions and future works.

2. THE SPEECH DATABASE AND THE ACOUSTIC SEGMENTS

For the manual segmentation we used the 200 phonetically balanced sentences proposed in [1], which were recorded by a 30-year-old male speaker from Rio de Janeiro with a university scholarship. The utterances were read with a speech rate of approximately 5.5 syllables per second with 0.55 of standard deviation. The slowest and highest speech rate observed were 4.12 and 7.05 syllables per seconds, respectively.

Table 1: List of 76 acoustic units used in the segmentation.

Sils.	15. [6~]	30. [ej~]	46. [je]	61. [k]
1. [sil]	16. [e]	31. [oj~]	47. [jo]	62.[kS]
2. [sp]	17. [E]	32. [w6~w]	48. [ju]	63. [f]
3. [lp]	18. [e~]	33. [Ej]	49. [wa]	64. [v]
4. [pl_p]	19. [i]	34. [Oj]	50. [we]	65. [l]
5. [pl_t]	20. [i~]	35. [Ew]	51. [wi]	66. [L]
6. [pl_k]	21. [o]	36. [Ow]	52. [wo]	67. [m]
7. [pl_kS]	22. [O]	37. [aw]	**Cons.**	68. [n]
8. [pl_pS]	23. [o~]	38. [ew]	53. [b]	69. [J]
9. [pl_tS]	24. [u]	39. [iw]	54. [d]	70. [R]
10.[pl_b]	25. [u~]	40. [ow]	55. [dZ]	71. [X]
11.[pl_d]	**Diphths.**	41. [aj]	56. [g]	72. [r]
12.[pl_dS]	26. [6~w]	42. [ej]	57. [p]	73. [s]
13.[pl_g]	27. [we~]	43. [oj]	58. [pS]	74. [z]
Vws.	28. [w6~]	44. [uj]	59. [t]	75. [S]
14. [a]	29. [uj~]	45. [ja]	60. [tS]	76. [Z]

Data segmentation was done using the PRAAT program [2]. To begin the segmentation we used a set of 37 phones proposed in [1]. However, this table was modified to characterize, with higher accuracy, the speech signal. We introduced, as acoustic units, silent intervals, vocalic transitions, vibrant and sibilant sounds. A sum of 39 acoustic units was added to the 37 phones initial table. The 76 acoustic units shown on Table 1 are divided into four sound categories: silent intervals (Sils.), vowels (Vws.), diphthongs/triphthongs (Diphths.) and consonants (Cons).

It is important to notice that diphthongs resulting from sandhi are included in the diphthongs category. We highlight also on Table 1 that, [sil], [sp] and [lp] stands for *silence*, *short pause* and *long pause*, respectively. The [*pl_phone*] units refer to the pauses before plosives. Each acoustic unit, except for the silences, was labeled according to the alternative version of the International Phonetic Alphabetic characters (IPA) proposed in [3, 4], that adapts the IPA symbols to the international keyboards. This symbol list, known as SAMPA, speeds the transcription and

ISBN 1-876346-48-5

Table 2: Length and primary stress

	S	PRE	POS
Phones Occ	34	30	19
DM (ms)	120.18	64.36	50.07
SD	46.18	29.68	42.13
Max. (ms)	335	190	217
Min. (ms)	23	0	0
Syll Occ	949	1056	635
D	0	66	153
D (%)	0	6.3	24.1

Table 3: Length and oral vowels

S	a	E	e	i	o	O	u
DM(ms)	124	110	109	95	115	134	106
SD	42.0	40.0	35.7	35.7	35.1	47.0	38.0
Syll Occ	241	93	71	101	40	52	23
D	-	-	-	-	-	-	-
PRE	a	E	e	i	o	O	u
DM(ms)	74	71	51	47	63	100	53
SD	21.4	0.5	33.6	22.9	27.4	0	22.0
Syll Occ	229	3	181	130	67	1	123
D	0	0	47	11	5	0	5
D (%)	0	0	26.0	8.5	7.5	0	7.5
POS	a	E	e	i	o	O	u
DM(ms)	59	-	65	50	119	-	47
SD	31.6	-	12.5	28.7	25.6	-	28.6
Syll Occ	188	-	68	77	63	-	85
D	13	-	66	12	60	-	2
D (%)	6.9	-	97.6	15.6	95.2	-	2.4

Table 4: Length and nasal vowels

S	6~	e~	i~	o~	u~
DM(ms)	117	104	90	123	85
SD	36.1	39.31	33.29	31.58	36.35
Syll Occ	23	36	41	24	24
PRE	6~	e~	i~	o~	u~
DM(ms)	68	72	63	58	63
SD	12.1	18.1	18.1	14.2	30
Syll Occ	14	30	37	30	39
POS	6~	e~	i~	o~	u~
DM(ms)	67	67	55	-	59
SD	31.6	5.2	12.9	-	38.6
Syll Occ	34	5	12	-	11

Table 5: Length and nasal diphthongs

S	6~w	ej~	oj~	uj~	w6~	we~
DM(ms)	140	116	141	113	80	83
SD	47.7	29.7	54.4	35.1	10.5	0
Syll Occ	53	38	7	14	4	1
PRE	6~w	ej~	oj~	uj~	w6~	we~
M	97	104	75	85	76	-
SD	44.5	29.7	20.5	10.5	13.2	-
Syll Occ	3	19	5	2	6	-
POS	6~w	ej~	oj~	uj~	w6~	we~
DM(ms)	111	93	-	-	77	-
SD	24.6	28.2	-	-	12	-
Syll Occ	5	8	-	-	2	-

facilitates the hard manual segmentation process. The boundaries of each segment were introduced according to four labeling levels: Phones, Syllables, Words and Phrasal Groups. In the first level of the segmentation, 6579 phones were labeled. In the second level, 3107 syllables were transcribed orthographically, even if some phones were not observed or if a reduction process happened. The lexical or primary stress was marked in the beginning of each tonic syllable.

In the third line of segmentation a group of 1780 Words was transcribed orthographically. Finally, for the last level of segmentation, an amount of 928 Phrasal Groups were determined considering the potential positions for the syntactic boundaries and observing the prosodic boundaries in the temporal pattern or in the speech melody. The phrasal or secondary stresses were placed in the beginning of the words that were phrasal topic or phrasal focus.

3. LENGTH ANALYSES OF VOCALIC SEGMENTS

In this section we present a study of the duration mean of the vocalic segments from the database used. It is analyzed the vowel length of the segmented units in function to the primary or lexical stress (3.1), the secondary or phrasal stress (3.2) and the position of the segments in the utterance (3.3). The length analyses of the 39 vocalic segments are described with respect to the: different number of phones in each syllable stress position (Phones Occ.); duration mean (DM); standard deviation (SD); maximum (Max.) and minimum (Min.) duration values of the segments; total syllable occurrences (Syll Occ); deletion (D) occurrences due to vowel reduction, as stated in [5] and the deletion percentage (D (%)) occurrence.

3.1 LENGTH AND PRIMARY STRESS

The vocalic segments observed are shared according to the stress status of the syllables with the following hierarchy positions: stressed (S), prestressed (PRE) and poststressed (POS).

For the S category, the duration mean of the stressed vocalic segments are considered in the lexical words (verbs, nouns, adjectives, adverbs) and in the grammatical words (prepositions, conjunctions, pronouns, numbers, possessives) with more than one syllable. In the PRE category, monosyllabic grammatical words, besides lexical words, are taken into account considering the duration of the prestressed vocalic segments. Finally, in the POS

category, only the lexical words are considered, regarding the duration of the poststressed vocalic segments. Table 2 shows the duration mean of the correspondent phones taking into account the stress status. The results shown on this table confirm the stress status of syllables in BP: the number of phones and duration mean of segments decrease as going from S to POS while the total deletion by vocalic reduction increases.

The duration means of each vocalic segment are detailed from Table 3 to Table 7.

From Table 3 it can be seen that the deletion on the oral vowels is not observed in the S position. In the PRE category, the phone [e] has the highest deletion percentage. Except for [a], [E] and [O], total reduction can appear in all other phones. The phones [E] and [O], observed in compound nouns, have a low occurrence number in this stress position. In the POS category, all phones occurred can have deletions. It is observed that almost all the phones [e] and [o] have total reduction.

Regarding Table 4, it can be seen that the duration means of the nasal vowels were shorter than the values related in [6-8]. Also, comparing the oral and nasal vowels a similar difference was noted. This discrepancy can be explained because, to obtain more accurate models, we considered the vocalic transition to the syllabic codas as part of the nasal consonants [m], [n] and [J]. With respect to the stress status of syllables, the highest duration difference is observed comparing the stressed and unstressed segments. The duration difference between PRE and POS nasals vowels is negligible.

Table 5 shows that the most frequent nasal diphthongs in S position are [6~w] and [ej~]. In PRE and POS categories the [ej~] occurrence is predominant. The diphthongs formed from the [6~] have mostly higher duration in POS than in PRE category.

Data from Table 6 shows that the falling oral diphthongs, listed on Table 1 (37-44), are the most frequent vocalic sequences. All these diphthongs occur in stressed position. The segments [iw] and [uj] do not occur in the PRE category. In the POS category only [ej], [ew], [ow] and [iw] occur with a duration mean slightly superior to the one in the PRE category. With respect to the rising oral diphthongs, listed on Table 1 (45-52), most of the occurrences are in the PRE category. In the S category, the vocalic sequences formed from [a] are the most frequent. Once more, the duration mean of the POS diphthongs overcome the PRE ones. The falling oral diphthongs, listed on Table 1 (33-36), occur only in the S category, except for [Ew], which is originated from vocalic sandhi and has a duration mean slightly longer than the S positioned [Ew].

3.2 LENGTH AND SECONDARY STRESS

To allocate the secondary stress, a perceptual experiment, with four listeners, was performed. The subjects analyzed and chose the proeminent words on each utterance. From the 1780 segmented words of the speech database, 325 were selected by at least 3 listeners. From this word selection a total of 318 stressed syllables, 267 prestressed syllables and 245 poststressed syllables were analyzed, as shown on Table 7.

Comparing Table 7 with Table 2, we verified that the vocalic lengthening that are the topic or in the phrasal focus occurs only in S position. The vocalic duration mean of the entire speech database is 120ms in S position while the duration mean for the syllables that have the primary and secondary stress have an increase of 10%, achieving 132ms. In the PRE and POS categories, the secondary stress does not change the duration mean of the syllables, as can be seen on Tables 7 and 2. With respect to the deletion, in the PRE category, the deletion percentage decreases from 6.3% to 2.6% while the similar vowel reduction for the POS category almost not changes, varying form 24.1% to 22.9%.

Table 6: Length of oral diphthongs

S	aj	aw	ej	ew	oj	ow	iw	uj
DM(ms)	154	164	129	117	134	168	145	115
SD	46	57	27	24	59	50	29	15
Syll Occ	26	28	28	14	19	26	3	4
S	ja	wa	jo	wi	Ej	Ew	Oj	Ow
DM(ms)	181	145	80	139	209	72	166	156
SD	54	57	0	6	89	22	39	20
Syll Occ	7	10	1	3	3	2	4	2
PRE	aj	aw	Ej	ew	oj	ow	ja	wa
DM(ms)	92	104	106	111	93	115	83	72
SD	21	39	12	25	37	0	27	22
Syll Occ	23	20	7	10	2	1	16	16
PRE	je	we	jo	wo	Ew			
DM(ms)	107	93	90.43	71	79			
SD	0	16	51.9	15	19			
Syll Occ	1	3	7	2	2			
POS	ej	ew	ow	iw	ja	wa	ju	
DM(ms)	108	128	139	120	95	87	97	
SD	0	9	0	74	45	40	51	
Syll Occ	1	2	1	2	38	9	24	

Table 7: Length and secondary stress

	S	PRE	POST
Phones Occ	29	23	16
DM(ms)	132	68	49
SD	43.0	26.4	38.9
Maximum	268	141	207
Minimum	77	0	0
Syll Occ	318	267	245
D	0	7	56
D (%)	0	2.6	22.9

ISBN 1-876346-48-5

Table 8: Length and Utterance Position

	S1	S2	U1	U2
DM	121	146	73	73
SD	40.2	44.6	26.5	57.3
Max	235	335	168	217
Min	23	31	zero	zero
Syll Occ	200	200	252	161
D	-	-	4	47
D (%)	-	-	1.6	29.2

3.3 LENGTH AND UTTERANCE POSITION

To verify the relationship between the initial and final length of syllables in the database utterances, we selected the duration of the first and last vocalic segment in stressed syllables, as seen on Table8, S1 and S2, respectively. We also selected the duration of the unstressed vocalic segments in the beginning of the utterances until the first stressed syllable, U1 from Table 8. From the last stressed syllable until the end of the utterances, we select the duration of the vocalic segments, U2 from Table 8.

The position of the syllable in the utterances is pertinent to determine the duration of the vocalic segments on the stressed syllable. The duration mean of the stressed syllable nucleus at the end of the utterances are 20% longer than the stressed syllable from the beginning of them. In the unstressed vocalic segments duration mean there is no difference between the PRE or POS category, as shown on Table 8. However, analyzing the standard deviation and the deletion percentage of the vocalic segments at the end of the utterance, there is more length variation in the POS position.

4. DISCUSSION

To establish the duration modeling for a PB speech synthesis system it is important to consider that the vocalic segment length increases in the stressed syllables. On the other hand, the relevance of the unstressed vocalic segment decreases in opposition to the consonant segments. These obtained results confirm what is stated in [8].

With respect to the secondary stress, it only affects the stressed vocalic units, although this is not the unique possibility. Pitch variations and the left displacement of the stress syllable are also another way to determine the secondary stress.

Regarding the utterance position, the duration of the final syllables are modified. A lengthening of the stressed syllables is observed, besides the behavior variance of the poststressed vocalic segments that can have even lengthening or deletion.

5. CONCLUSIONS

This paper shows the units table that directed the segmentation and provides relevant data to obtain a duration modeling for the Brazilian Portuguese (BP) spoken in Rio de Janeiro. The stressed position, the speech rate, the secondary stressed distribution as well as the syllabic position at the beginning or end of the utterance are relevant factors to establish the acoustic units' duration mean.

Future works enclose consonantal behavior research that could complement the study of the vocalic segments presented in this paper, especially to verify the consonant growing rate in unstressed position related to the vowel reduction. Applying the duration modeling to the TTS system for BP is also a future work.

REFERENCES

[1] A. Alcaim, J. A. Solewicz and J. A. de Moraes, "Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro", *Revista da Sociedade Brasileira de Telecomunicações*, vol. 7(1), pp. 23-41, 1992.

[2] P. Boersma and D. Weenick, "Praat: doing phonetics by computer," Institute of Phonetic Sciences, University of Amsterdam, <www.praat.org>, 2003.

[3] Speech Assessment Methods Phonetic Alphabet (SAMPA).Available in <www.phon.ucl.ac.uk/home/sampa/home.htm >. Accessed in: 03/10/2003.

[4] J. Wells, "Computer-coding the IPA: a proposed extension of SAMPA," Department of Phonetics and Linguistics, University College of London, <http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>, 27/06/2002.

[5] R. M. Dauer, "Stress-timing and syllable-timing re-analyzed," *Journal of Phonetics*, vol. 11, pp. 51-62, 1983.

[6] P. Barbosa, "Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia de fala", in *Estudos de prosódia*, E. Scarpa, Ed., São Paulo, UNICAMP, pp. 21-52, 1999.

[7] J. A. de MORAES, "Um algoritmo para a correção/simulação da duração dos segmentos vocálicos em português", in *Estudos de prosódia*, E. Scarpa, Ed., São Paulo, UNICAMP, pp.69-81, 1999.

[8] C. Reis, *L'Intéraction entre l'Accent, l'Intonation et le Rythme en Portugais brésilien: étude acoustique de la prosodie,* Thèse de Doctorat, Université de Provence 1, Institut de Phonétique, Aix-en –Provence, 1995.