Cad.Est.Ling., Campinas, (31) :33-53, Jul./Dez. 1996

AT LEAST TWO MACRORHYTHMIC UNITS ARE NECESSARY FOR MODELING BRAZILIAN PORTUGUESE DURATION: EMPHASIS ON AUTOMATIC SEGMENTAL DURATION GENERATION [*]

PLÍNIO ALMEIDA BARBOSA
(UNICAMP)

> *Et le vent, la vague, l'étoile, l'oiseau, l'horloge, vous répondront, il est l'heure de s'enivrer ;* pour ne pas être les esclaves martyrisés du temps, *enivrez-vous, enivrez-vous sans cesse de vin, de poésie, de vertu, à votre guise.*
> Les Petits poèmes en prose, *Charles Baudelaire*

ABSTRACT: By modeling Brazilian Portuguese acoustic duration, this work presents two arguments in favor of macrorhythmic units. First, the emergence of distinct durational patterns for lexical and phrasal accents. Second, the homogeneous lengthening (shortening) effect of segments correlating syllables at lexical stress and IPCGs at phrasal accent. A two-stage model of segmental duration generation is derived.

RÉSUMÉ: La caractérisation de la durée en portugais du Brésil (PB) permet de faire émerger une typologie accentuelle signalant la présence de deux unités macrorythmiques. Les maxima des *z-scores* de la syllabe coïncident avec la position de l'accent lexical tandis que les maxima des *z-scores* du GIPC démarquent les frontières prosodiques de l'énoncé. Un modèle à deux étapes permettant la génération simplifiée de la durée segmentale du PB en est dérivé.

RESUMO: O modelamento da duração acústica no português do Brasil (PB) tornou possível a emergência de uma tipologia acentual que revela a existência de ao menos duas unidades de programação macrorrítmicas: a sílaba e o GIPC. Os pontos de máximo dos *z-scores* da sílaba coincidem com a posição do acento lexical enquanto os

[*]A shorter version of this paper is published on the *Proceedings of the 1st ESCA Tutorial and Research Workshop on Speech Production Modeling & 4th Speech Production Seminar*, Autrans, France, May 20 to 24th, 1996, under the title: *At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration.*

pontos de máximo dos *z-scores* do GIPC (coincidindo com posição de acento lexical) demarcam as fronteiras prosódicas do enunciado. Um modelo rítmico possibilitando a geração automática e simplificada da duração segmental é proposto para ser integrado em um sistema de síntese da fala em PB.

INTRODUCTION

Recently, van Santen's work (1994) convincingly showed that it is not necessary to consider the existence of (macro)rhythmic programming units (RPU) in order to generate segmental duration. His phoneme-based approach involves, however, a huge computational cost. In early work, on other hand, Barbosa and Bailly (1997) have shown that the durational structure (expressed by a syllable size z-score model) associated with a set of read sentences reveals a certain kind of organization over the segment level. We were able to propose an approach for duration generation that takes into account the macrorhythmic organization of speech and drastically simplifies the mechanism of duration assignment.

As has already been demonstrated for French, normalizing the acoustic duration of consecutive segments points to an organization into higher order units whose boundaries are two consecutive vowel onsets (Barbosa & Bailly 1994). This unit was named inter-perceptual-center group (IPCG) by reference to research on p-centers (Marcus 1981; Pompino-Marschall 1991), whose findings suggest that their optimal location is the vicinity of the vowel onset, as can be observed by careful examination of Pompino-Marshall's figures and can be confirmed in a more ecological work (Janker 1995).

Normalized durations are obtained through Campbell's *z-score* model (1992). The z value of each segment *s* is computed by writing:

$$Dur_s = \exp(\mu_s + z.\sigma_s) \qquad (1)$$

where Dur_s is the segment duration and (μ_s, σ_s) stands for the average and the standard-deviation of the log-transformed durations of all *s* realizations in an *ad hoc corpus*. The strong elasticity hypothesis in Campbell's model says that all segments in a syllable frame have the same z-score: a single value of *z* per syllable can then be computed by writing:

$$Dur(\text{syllable}) = \sum_s \exp(\mu_s + z.\sigma_s) \qquad (2)$$

French data allow us to propose a weaker elasticity hypothesis where the rhythmic unit is the IPCG, not the syllable. Brazilian Portuguese (BP) data, on the other hand, indicate that at least two macrorhythmic units are necessary to model the durational structure of read sentences. BP rhythm can then shed light on the segmental/suprasegmental controversy thanks to the greater complexity of its accentual typology.

In BP, lexical stress can be assigned to the final, penultimate or antepenultimate syllable. Stressed syllables can be enhanced as they are uttered by carrying phrasal accent. Only lexically stressable syllables can bear phrasal accent.

The acoustical correlates of stress are often the greater duration of the stressed unit and the decrease of intensity in the post-stressed syllables (if any) (Massini 1991)[1]. Massini also says that stress is carried by the syllable as a whole – and not by the vowel alone– but our data reveal a slightly more complex situation.

The results shown below are based on the analysis of two *corpora*. A nonsense word *corpus*, meant to capture the speaker durational characteristics, and a read-sentence corpus, meant to analyze utterances' rhythmic structure.

SPEAKER'S DURATION STATISTICS

Segmental durations were determined from a 1195-nonsense word *corpus*, containing all BP phonemes (and also some of the BP allophones) of a native 30-year-old professional speaker (from the Paulista dialect). Statistical analyses were then performed. The results on Table 1 confirm current knowledge on duration in BP (which is in agreement with universal trends: Lehiste 1970): (a) for front and back vowels, the higher the vowel, the shorter its average duration; (b) post-stressed vowels (/ɐ, ɪ, ʊ/) are shorter than their stressed counterparts (/a, i, u/); (c) nasal vowels are longer than their oral counterparts; (d) voiceless consonants are longer than their voiced counterparts.

Table 1: Mean duration (and standard-deviation) of the BP phones (in ms) for our speaker[2].

i	145 (37)	ɐ	111 (45)	ĩ	209 (25)	tʃ	149 (20)	f	138 (14)	n	76 (15)
e	170 (36)	ɪ	98 (44)	õ	229 (26)	k	121 (21)	s	143 (26)	ɲ	103 (24)
ɛ	175 (32)	ʊ	77 (19)	ũ	215 (29)	b	86 (17)	ʃ	143 (16)	ɾ	47 (16)
a	165 (28)	j	92 (10)	j̃	136 (14)	d	71 (17)	v	78 (16)	r	81 (12)
u	134 (42)	w	97 (25)	w̃	139 (23)	dʒ	109 (18)	z	87 (21)	ʁ	62 (15)
o	168 (35)	ɐ̃	174 (46)	p	120 (20)	g	67 (16)	ʒ	89 (12)	l	73 (16)
ɔ	183 (29)	ẽ	210 (44)	t	113 (20)			ɾ̃	90 (12)	ʎ	77 (14)

The log-transformed versions of these data were used in formula (2) (where *syllable* is either phonological syllable or IPCG) to compute the z-scores of syllables and IPCGs of 100 sentences read by the same speaker. This corpus was manually segmented and carefully labeled by the author (a total of 2,055 syllables). Sentence length varies between one and 84 syllables. Syntactic boundaries were also marked using a set of eight hierarchical labels (obtained by the projection of a surface tree – from a

[1]Fundamental frequency is not an acoustic correlate of lexical stress but is a cue of phrasal prominence.

[2]The /r/ phoneme was realized as a trill ([r]) in syllable-final position and as a fricative (transcribed as the uvular [ʁ]) elsewhere (e.g. in *carro* or *rosa*).

dependence grammar where the root is the verb – over the paradigmatic axis. See Barbosa & Bailly 1994 for a more detailed description of these labels).

EXTRACTING DURATIONAL CONTOURS FROM Z-SCORE EVOLUTION

Segments were grouped into two kinds of RPU's: syllable and IPCG. By using the raw duration of each group in formula (2) above, the *z-scores* were computed for all RPU's in each sentence. An example is shown in Figure 1.

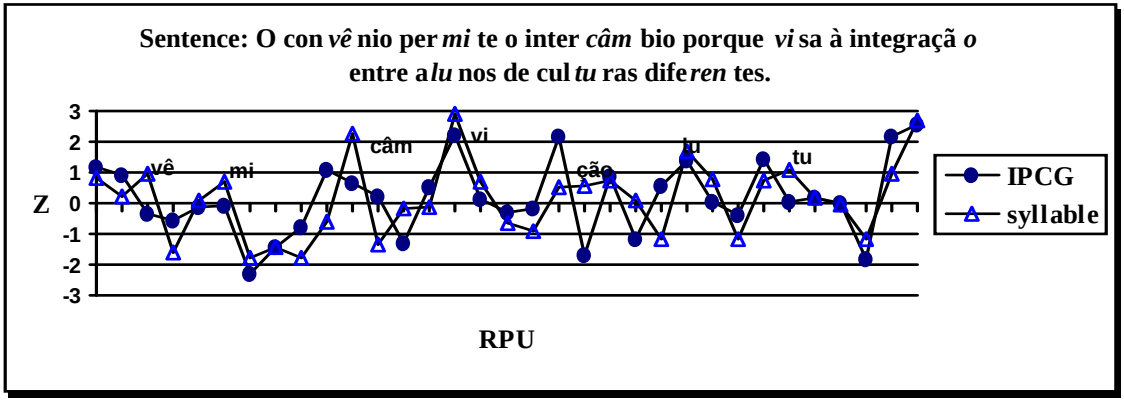

Figure 1: Z-scores (vertical axis) for IPCG's and syllables in the sentence "O convênio permite o intercâmbio porque visa à integração entre alunos de culturas diferentes." ("The agreement allows for interchange since it aims at the integration among students from different cultures."). Lexically stressed syllables are italicized. Tick marks on the horizontal axis represent the vowels along the sentence (signaling both syllables and IPCG's; NB: orthographic *-io* – in *convênio* and *intercâmbio* – is a semivowel/vowel cluster). Note that the greatest z-scores for the syllable correspond to lexically stressed vowels.

In all 100 rhythmic contours, syllable z-scores indicate the (lexically) stressed syllables of the utterance: the highest z-score within each word coincides with the lexically stressed syllable. On the other hand, if the highest IPCG z-scores within non-clitic words *at lexical stressed position* are taken as a criterion for boundary placement and as a measure of boundary strength, coherent prosodic groups are obtained for all sentences in the *corpus*. In Figure 1, the strongest boundary (IPCG z-score of *vi* in *visa*) splits the sentence into two chunks of 16 IPCG's each. The property of eurhythmicity is very clear and the coherence between this result and those of Grosjean's performance trees (1993) is notorious.

IPCG z-scores delimitate accentual groups (prosodic words) where rhythmic pattern is characterized by frequent alternation of z-score values at the beginning of the accentual group followed by a duration *crescendo* (starting at least on the penultimate syllable) towards the last stressed syllable in the group.

STATISTICAL CONFIRMATION OF THE RESULTS

Statistical analyses confirm that actual segmental z-scores (by using raw segmental duration into formula 1) are strongly correlated in RPU frames. For these analyses, phrasal boundaries were marked by choosing positions in the utterance corresponding to IPCG z-score maxima (coinciding with lexically stressed RPU) and by carefully hearing the utterances in order to confirm these candidates to prominence.

The results in table 2 show that, in phrasal accent position, *onset/nucleus* segment sequences (in the syllable) are negatively correlated (-31%) whereas *nucleus/coda* segment sequences (necessarily in the rhyme) are positively correlated (76%). On the other hand, *onset/ nucleus* and *nucleus /coda* segment sequences confirm that lexically stressed syllables not bearing phrasal accent form a homogeneous unit. In non prominent positions, VC and V#C sequences seem to suggest that the IPCG is a homogeneous unit whose rhyme component is enhanced at phrasal accent position.

Table 2: Correlation (in percentage) between consecutive segmental *z-scores* according to accentuation degree. Segments were categorized for each phonological syllable with three labels: *onset*, for each segment in onset position, *nucleus*, for the vowel nucleus, and *coda*, for each segment in coda position. Sequences as *nucleus-onset* span necessarily over syllable boundaries. The lexical stress category refers to lexical stressed RPU not bearing phrasal accent. Only the significant values are reproduced here.

	lexical stress	***phrasal accent***	***other positions***
onset/nucleus	63	-31	4
nucleus /onset	ns	26	56
nucleus /coda	48	76	63

These results corroborate Campbell's predictions (1993) only in part: since he adopts the syllable as the RPU, onset segments in syllables at prosodic boundaries are overpredicted. The observed final lengthening affects the rhyme, not the onset. This is produced by lesser articulator stiffness associated with closure movement (Edwards, Beckman, & Fletcher 1991).

NEURAL NETWORK ARCHITECTURE, TRAINING AND TEST

RPU z-scores can be a means of deriving the segmental durations of a particular sentence (if formula 1 is applied by setting $z = z_{syllabe}$ or $z = z_{IPCG}$). The above durational contours can be easily generated by neural networks, as implemented by the author for French (Barbosa & Bailly, *op.cit.*). In this work segmental durations were computed in a second stage by sequentially applying formula 2 and 1 (these two steps constitute the so-called *repartition algorithm*) with the IPCG duration delivered by the network output (this two-stage model is proposed in Campbell 1992).

For French, a sequential, recurrent network was used to learn to associate a phonological, prosodic description of the sentence to the respective rhythmic pattern expressed by the evolution of IPCG duration over the sentence. This method allows for preservation of macrorhythmic unit durations maintaining the rhythmicity of the original sentence. Although the vowel quality of the syllable nucleus is sequentially described at the network input, only the number of the intervening consonants between two vowels and not their nature are represented. This implicitly means that consonant nature is not important to derive IPCG duration (although microrhythmic differences in segmental duration can be captured in the second stage by the repartition algorithm).

For BP, a simpler network was used by implementing a multilayered perceptron (Rohani, Chen & Manry 1992). At the perceptron input, a phonological, prosodic description of each sentence is used to infer the network output, the IPCG and syllable z-score evolution over the sentence.

Thanks to a greater coherence between accentual typology and z-score patterning, the network learning was, in fact, faster. Furthermore, z-score patterns are smoother than that of RPU duration. But original RPU durations are no longer preserved. What is preserved in this framework is the rhythmic structure as represented by the z-score patterning. In a typical consonant contrast as *ti/bi*, for instance, different consonants (/t/ and /b/) induce different RPU durations, since the same RPU z-score delivered by the network allows to obtain different consonant durations in the repartition algorithm but the same vowel duration for /i/. In the early French version an identical duration for the pair *ti/bi* would have been imposed and different durations for /t/, /b/ and /i/ would have been obtained.

Our model of segmental duration generation was applied to the learning and also to the test *corpus* subsets. The model was capable to generalize even when lexical stress position was manipulated.

The 17 formal neurons used at the perceptron input stand for a phonological, syntactic description of each sentence. They are coded with real numbers between 0 and 1 and are briefly described below.

Internal clock. This is a measure of the utterance speech rate. This value is computed per sentence by averaging the IPCG durations not associated with a syntactic marker[3];

Declination line. (Decreasing) number of IPCGs between the current position and the final in the sentence. Normalization factor: 100;

Lexical stress (pre-stressed, stressed and post-stressed) 3 IPCGs before the current position;

Lexical stress (pre-stressed, stressed and post-stressed) 2 IPCGs before the current position;

Lexical stress (pre-stressed, stressed and post-stressed) immediately before the current position;

Lexical stress (pre-stressed, stressed and post-stressed) at the current position;

"Declination line" associated with the current phrase. (Decreasing) number of IPCGs between the current position and the next syntactic marker;

Syntactic marker that dominates the current phrase;

Syntactic marker that dominates the next phrase;

Vowel nature 3 GIPCS before the one at the current position;

Vowel nature 2 GIPCS before the one at the current position;

Vowel nature immediately before the one at the current position;

Vowel nature at the current position;

Number of consonants in the IPCG immediately before the current one;

Number of consonants in the current IPCG;

Number of consonants in the coda of the syllable immediately before the current one;

Number of consonants in the coda of the current syllable.

[3]The syntactic boundaries were marked manually (an automatic assignment can be obtained with a parser). A set of nine distinct markers were extracted from a dependence grammar analysis where the head is the verb (Tesnière 1965, Martin 1981). This set is a modified version of Bailly's markers (Barbosa & Bailly 1994). These markers are obtained by projecting the surface tree nodes over the syntagmatic axis. The strength between adjacent nodes is indicated by the dependence relation between them. The markers are: IF (the two adjacent nodes come from distinct trees which is common with strong ponctuation signs and coordinated clauses); TF (the two adjacent nodes are dominated by the verb); DF (the dominated node is to the right of the dominant verb. Example: between verb and complement); GF (the dominated node is to the left of the dominant verb. Example: between subject and adjacent verb); ID (the two adjacent nodes are not directly related but are in the same tree); DD (the dominated node is to the right of the dominant one, which is normally a noun. Example: between noun and postposed adjective); DG (the dominated node is to the left of the dominant one, which is normally a noun. Example: between anteposed adjective and noun); IT (the two adjacent nodes are dominated by a same node which is not the verb); FF (sentence ending). Here two examples: "O gatinho <GF> bebeu <DF> leite <TF> numa tigela <DD> verde <FF>." e "Ontem, <IF> o calmo <DG> gatinho <DD> preto <ID> bebeu <DF> leite <TF> numa tigela <DD> verde <IT> e rosa <FF>."

The number of (formal) neurons in the hidden layer (25) was estimated by Manry's software (for a global learning error of 0.5. This error is computed between the desired IPCG and syllable z-scores - previously calculated using formula 1 - and the IPCG and syllable z-scores obtained at the network output). The 2 nodes at the output stand for the IPCG and syllable z-score corresponding to a current position in the sentence (current syllable and IPCG sharing the same vowel).

During the learning phase, 39 sentences were included in the training (these sentences constitute the learning *subcorpus*. The remaining ones, the test *subcorpus*). They contain 663 RPUs presented iteratively to the network by examples. An example is a pair formed by the linguistic description of the current RPU (performed by the 17 nodes described above) at the input and the corresponding IPCG and syllable z-scores at the output.

The network computes at each iteration the error between the original and the estimated z-scores (for IPCGs and syllables). This error allows to modify the connections' weights in the direction of a better approximation between estimated and original outputs.

The degree of convergence for this process is satisfactory, as can be seen by the gradation of estimated rhythmic patterns modifications in the sequence of figures 3, 4 and 5 (for IPCGs only). The sentence in these examples is: "As taxas de juros no mercado interno estão subindo bastante."

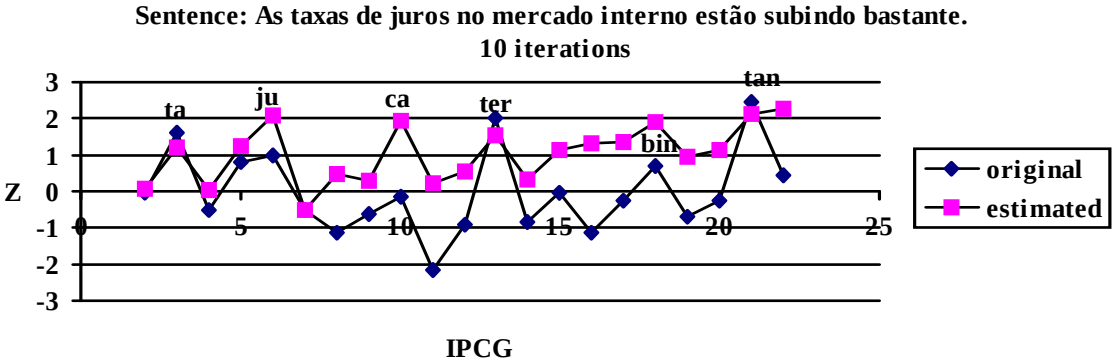

Figure 3: Comparison between original IPCG z-scores (presented to the network as the output element of this example) and IPCG z-scores estimated by the network after 10 iterations. Notice the distance between original and estimated rhythmic patterns. After 10 iterations the phrasal accent at "-ter" (from "interno") is already reproduced. Syllable (instead of IPCG) orthography is indicated at corresponding positions for ease of reading.

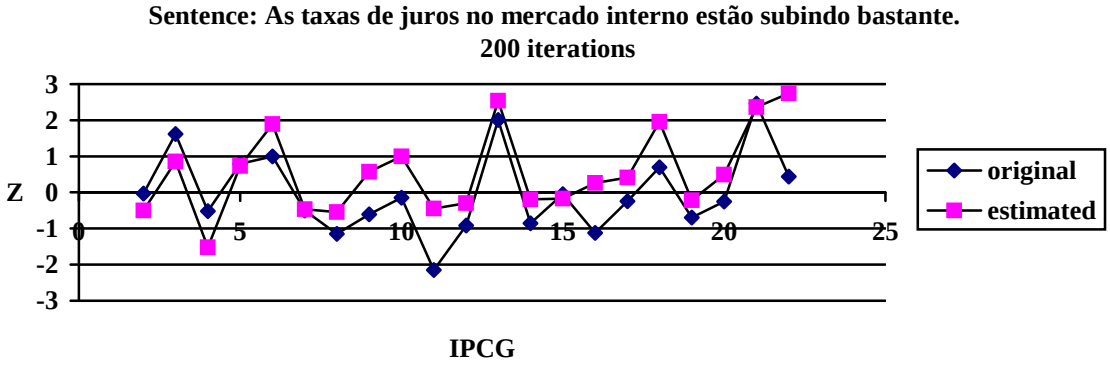

Figure 4: Comparison between original IPCG z-scores (presented to the network as the output element of this example) and IPCG z-scores estimated by the network after 200 iterations. Notice that the pattern corresponding to “interno” was improved compared to the one in the previous figure. In 200 iterations the sentence rhythmic pattern is well reproduced.

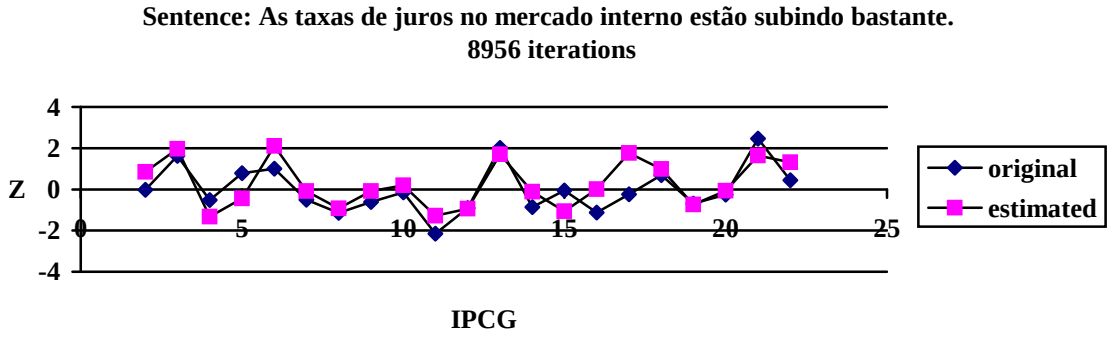

Figure 5: Comparison between original IPCG z-scores (presented to the network as the output element of this example) and IPCG z-scores estimated by the network after 8956 iterations. Although the pattern is not perfectly well reproduced, the network captured the saliences of "ju-" (from "juros", 5th position) and "subindo" (16th, 17th and 18th positions).

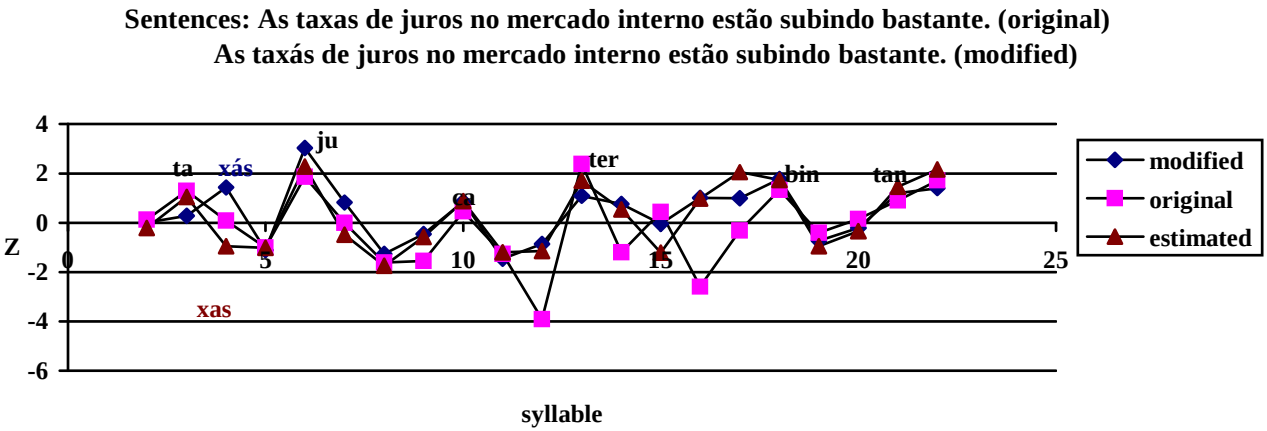

Figure 6: Comparison between the rhythmic patterns for the syllables in the sentence “As taxas de juros no mercado interno estão subindo bastante.” (original) and “As taxás de juros no mercado interno estão subindo bastante.” (modified). Notice that the accentual peak (learned by the network) on “ta-” (2th position) migrates to the 3th position (corresponding to “xás”). The next z-scores are maintained close to each other (compare remaining positions for estimated and modified z-scores).

The evolution of the syllable z-scores has the same degree of coherence and can be seen in figure 6 (compare original and estimated z-scores). As it is shown here, the network is capable of learning the sentence rhythmic patterns (actual *gestalten*) by estimating from the input what is more important to determine the desired output. It can be said that the neural net is a data-driven system because it learns from the empirical domain represented by the speech *corpora*.

Generalizing still more means to be able to capture the underlying mechanism allowing to associate input and output, that is, to learn to pacemake from a symbolic representation. In order to test this capacity, a modification from the original sentence "As taxas de juros no mercado interno estão subindo bastante." is introduced here. The lexical stress on "ta-" (from "taxas") is placed at the next syllable ("-xas"), creating the pseudo word "taxás". The results are shown below.

Figure 6 shows that the neural net is able to mimic a rhythmic pattern very well. Nevertheless, the non exact matching between original and estimated patterns introduces additional errors to the segmental duration generation model when the network works together with the repartition algorithm.

SEGMENTAL DURATION GENERATION

A program in C language was developed to integrate the two stages of duration prediction (perceptron and repartition algorithm).

All segmental durations in the learning and test *subcorpora* were predicted by the model. It is also important to note that the estimated z-scores can generate a segmental duration (by using formula 2) smaller than the minimum allowed by the phonatory system. In order to prevent this serious error a minimum duration associated with each segment is computed from the corpora and used as a threshold. No segmental duration is generated for a particular phone below that minimum.

The error means between original and estimated duration were -1 ms for the learning subcorpus and 2 ms for the test subcorpus (both are not statistically different from 0). The standard-deviations are 32 ms, for the learning *subcorpus* and 36 ms, for the test one. As was said before, the standard-deviations are greater than that obtained at the output of the repartition algorithm alone. The closeness between standard-deviations for the two *subcorpora* is a sign of a good performance in generalizing.

Segmental durations predicted by the model are presented below for two sentences (in table 3, a learning *subcorpus* utterance and, in table 4, a test *subcorpus* one).

Table 3: Natural and estimated segmental durations (in ms). Estimation performed by the Barbosa-Bailly model with the sentence "As taxas de juros no mercado interno estão subindo bastante.", one of the sentences in the learning *subcorpus*.

Segment	a	s	t	a	ʃ	ɐ	z	dʒ	ɪ	ʒ	u	ɾ
Original duration	75	104	69	137	84	70	47	41	50	85	118	25
Estimated duration	108	97	78	108	71	40	38	53	46	92	147	59
Segment	ʊ	z	n	ʊ	m	e	r	k	a	d	ʊ	
Original duration	58	44	28	39	44	64	45	83	99	39	28	
Estimated duration	46	51	47	38	49	88	47	82	107	31	33	
Segment	ĩ	t	ɛ	r	n	ʊ	ɪ	s	t	ɐ̃	w̃	s
Original duration	75	86	175	54	35	38	68	86	46	51	49	106
Estimated duration	105	57	133	51	63	46	53	81	79	124	93	114
Segment	u	b	ĩ	d	ʊ	b	a	s	t	ɐ̃	tʃ	ɪ
Original duration	75	86	175	54	35	38	68	86	46	51	49	106
Estimated duration	105	57	133	51	63	46	53	81	79	124	93	114

Table 4: Natural and estimated segmental durations (in ms). Estimation performed by the Barbosa-Bailly model with the sentence "As operações de crédito continuam.", one of the sentences in the learning *subcorpus*.

Segment	a	s	o	p	e	ɾ	a	s	õ	j̃	z	dʒ	ɪ		
Original duration	64	62	70	81	71	16	78	140	71	77	35	56	24		
Estimated duration	75	38	79	60	69	17	59	59	105	66	34	85	118		
Segment	k	r	ɛ	dʒ	ɪ	t	ʊ	k	o	tʃ	ɪ	n	u	ɐ̃	w̃
Original duration	126	52	175	70	44	78	38	78	99	85	24	44	163	100	34
Estimated duration	99	54	117	74	74	76	45	69	103	66	55	47	126	102	83

EVALUATION OF THE RHYTHM GENERATION MODEL

The results presented here show clearly how it is possible to obtain a segmental duration generator integrating two important characteristics for a speech synthesis system: automation and correct reproduction of the BP natural rhythm. The capacity of the model to generalize to new sentences was also shown.

The network learning can be continued. A typical computation lasts 30 minutes for about 1,000 iterations. Nevertheless, for this generator, the performance of the entire model will never be better than the performance of the repartition algorithm alone (standard-deviation of 24 ms). But the results are very satisfactory. We cannot forget that a experimented phonetician makes mistakes with about 10-ms standard-deviations when segmenting the speech signal (Leung & Zue, 1984).

A perception test was also performed in order to evaluate the model of segmental duration generation. An ABBA test allowed us to evaluate 10 utterances whose segmental durations were modified[4] by analysis-resynthesis with the Hybrid Model (Böeffard & Violaro 1994). Segmental durations were assigned to utterances presented in pairs according to two models: our rhythmic model of segmental duration generation (utterances of type *model*) and a model with the same error distribution[5] as our generator but having durations assigned according to a Gaussian number generator (utterances of type *random)*. What is being evaluated when these two models are compared is the tendency for our rhythmic generator to preserve homogeneous lengthening of syllables at lexical stress and of IPCGs at phrasal accent. This tendency is not taken into account by the random model.

The ten pairs of utterances were presented to fifteen listeners. Each pair has a model utterance and a random utterance ramdomly ordered. Utterance pairs are also randomly organized in a sequence for each listener. During the session, each pair is heard twice by the listener via headphones (in this case, in the same order). After listening, the subject must decide which utterance seemed less unnatural (the first or the second one) by writing down on a specific sheet. A *don't-care* option could be used if the two utterances seemed identical. Some listeners' comments were taken at the session end.

The results point to a preference of about 67% (significantly different from chance) for the utterance modified by our segmental duration generator. All subjects said that the utterances sound quite artificial. (This aspect is inherent to the Hybrid Model, which is in phase of improvement.)

This weak - but stable - preference for our model can be explained by the type of test prepared. Both models have a 27-ms standard-deviation for segmental durations. If the perception thresholds for durations (30 ms for vowels and 40 ms for consonants) proposed by some authors like Goedemans & van Heuven (1995) are taken as true, an important amount of the utterances' duration errors would be very close or under the threshold. If this assumption is true, it is very hard for the listeners to perceive any

[4]Fundamental frequency and intensity were unchanged.

[5]Our generator presented a gaussian-like distribution of errors when comparing original and predicted durations. The two models have same means and standard-deviations.

difference between the two versions of the original utterance. A great amount of *don't care* options reinforces this hypothesis.

CONCLUSIONS

The systematic findings correlating lexical stress in syllable frames and phrasal accent in IPCG frames constitute arguments in favor of the existence of rhythmic units above the segment level.

These distinct macrorhythmic units are also an evidence for a lexical and a post-lexical rhythm component, as suggested by Keating (1995), and point to the need for a rhythm tier where rhythmic nodes dominate different linguistic units in explicit models of speech production (as outlined by Articulatory Phonology in Browman & Goldstein 1990). (This is equivalent to saying that models of speech production should receive explicit linguistic input. And, from this point of view, this work insists on the need for cognitive theories of phonetics.)

The persistence of rhythmic homogeneity of lexically stressed syllables at the utterance level may be explained by a strong constraint on phasing of articulatory gestures (Browman & Goldstein 1986; 1990; 1992) that operates on lexical units (a microrhythmic adjustment in order to enhance the stressed syllable of lexical entries, as suggested by Perkell 1980) and broadly maintains the former phase relations at the utterance level.

Post-lexical metrical (macrorhythmic) readjustments as the Rhythm rule and microrhythmic readjustments as external sandhi rules (very common in BP) manipulating the gesture constellation would intervene during the elaboration of the connected-speech plan for each utterance.

Finally, teleological production (and perception)-guided readjustments may also intervene. Their function has a connection with the need for monitoring utterance production and for ensuring that articulatory movements are produced comfortably (low in articulatory cost). (A connection with the need for ensuring listeners comfortable decoding would also have an important communicative function.) Research insisting on the existence of a internal[6] clock enabling and monitoring rhythmic productions such as speech may favor the former assumption (Turvey, Schmidt & Rosenblum 1990; Pöppel 1989; Semjen, Schulze & Vorberg 1992; Leiner, Leiner & Dow 1991). Research as the one carried out by Edwards, Beckman & Fletcher, 1991, showing that increased duration at phrasal final position is realized by stiffness decreasing of VC units confirms the latter assumption and the IPCG homogeneity at the phrasal accent level.

The comfortable speech production would be realized by ballistic closure (jaw) movements contained in the VC frame (the jaw O-C movement is the articulatory correlate of macrorhythmic unit succession). In this sequencing, vowels are the most important segments as it is pointed out by B&G (1990, p. 352): "The X-ray data we have analyzed (...) have consistently supported the contention that consonant articulations are superimposed on continuous vowel articulations, which themselves

[6]The word *internal* is used here in cognitive terms and means "in the brain".

minimally overlap." (In our work this is confirmed by the closeness of the vowel z-score contours with the IPCG ones. Syllables z-score contours are never correlates of phrasal accentuation.)

Under these assumptions, IPCGs and syllables have distinct status: the former are ease-of-production-and-monitoring units (or teleological units) at the phrasal level and the latter, gestural units (in the sense of the lexical orientation of Articulatory Phonology) at the word level. In the utterance framework, accentual groups are delimited by IPCG duration increasing followed by a durational contour reset (a typical alternation-*crescendo* contour. Whether this duration *crescendo* is related to a stiffness *descendo*, is a matter of investigation. But see Edwards, Beckman & Fletcher, 1991, note 2). Accentual group (AG) boundaries should also delimit regions where phenomena as sandhi rules or V-to-V coarticulation would be possible (but not across the AG boundaries).

ACKNOWLEDGMENTS

The author is currently supported by a grant (96/7832-4 and 95/9532-6) from the Fundação de Amparo à Pesquisa do Estado de São Paulo-FAPESP. The speaker and the listeners are strongly acknowledged. I am also in debt with Eleonora Albano e Sandra Madureira for their helpful suggestions in previous versions of this paper.

REFERENCES

BARBOSA, P.A. & Bailly, G. *Generating pauses within the z-score model*. In: *Progress in Speech Synthesis*. van Santen, J.P.H., Sproat, R.W., Olive, J.P. & Hirschberg, J. (Eds.), New York: Springer-Verlag, 1997:365-381.

BARBOSA, P.A. & Bailly, G. *Characterisation of rhythmic patterns for text-to-speech synthesis*, Speech Communication, 15 (1-2), 1994:127-137.

BÖEFFARD, O. & Violaro, F. *Using a hybrid model in a Text-to Speech system to enlarge prosodic modifications.* International Conference on Spoken Language Processing (ICSLP '94), Yokohama, Japan, 1994:727-730.

BROWMAN, C.P. & Goldstein, L. *Articulatory Phonology: an overview*. Phonetica, 49, 1992:155-180.

BROWMAN, C.P. & Goldstein, L. *Tiers in articulatory phonology with some implication for casual speech.* Kingston, J. & Beckman, M.E. (Eds.). Papers in Laboratory Phonology 1. Cambridge University Press, 1990:341-376.

BROWMAN, C.P. & Goldstein, L. *Towards an articulatory phonology.* Phonology Yearbook, 3, 1986:219-252.

CAMPBELL, N.W. Syllable-based segmental duration. In: Talking Machines: theories, models, and designs (Bailly, G. & Benoît, C. Eds.), 1992:211-224.

CAMPBELL, N.W. *Automatic detection of prosodic boundaries in speech*, Speech Communication, 13, 1993:343-354.

EDWARDS, J., Beckman, M.E. & Fletcher, J. *The articulatory kinematics of final lengthening*. J. Acoust. Soc. Am. 89 (1), 1991:369-382.

GOEDEMANS, R. & van Heuven, V.J. *Duration perception in subsyllabic constituents.* Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH), September, 18-21st, Madrid, Spain, 2, 1995:1315-1318.

JANKER, P.M. *On the influence of the internal structure of a syllable on the p-center perception.* XIII International Congress of Phonetic Sciences, August 13-19, Stockolm, Sweden, 2, 1995:510-513.

KEATING, P.A. *Segmental Phonology and Non-segmental Phonetics.* XIII International Congress of Phonetic Sciences, August 13-19, Stockolm, Sweden, 3, 1995:26-32.

LEHISTE, I. *Suprasegmentals.* Cambridge, Massachussets: MIT Press, 1970.

LEINER, H.C., Leiner, A.L. & Dow, R.-S. *The human cerebro-cerebellar system: its computing, cognitive, and language skills.* Behavioural Brain Research, 44, 1991:113-128.

LEUNG, H.C. & Zue, V. W. *A procedure for automatic alignment of phonetic transcriptions with continuous speech*. Proceedings of the IEEE ICASSP, 1, San Diego, 1984:2.7.1-2.7.4.

MARCUS, S.M. *Acoustic determinants of Perceptual-center (p-center) location*, Perception and Psychophysics, 30(3), 1981:247-256.

MARTIN, P. *L'Intonation est-elle une structure congruente à la syntaxe ?* In: *L'Intonation : de l'acoustique à la sémantique*, Paris: Klincksieck, 1981:234-271.

MASSINI, G. *A Duração no estudo do acento e do ritmo em português*, Master's thesis, Unicamp, 1991.

MONNIN & Grosjean, *Les Structures de performance en français : caractérisation et prédiction*, L'Année Psychologique 93, 1993:9-30.

PERKELL, J.S. *Phonetic Features and the Physiology of Speech Production.* In: *Language Production: Speech and Talk.* Butterworth, B. (Ed.). Academic Press: London. Vol 1, 1980: pp 33-372.

POMPINO-MARSCHALL, B. *The syllable as a prosodic unit and the so-called P-centre effect*. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, 29, 1991:65-123.

PÖPPEL, E. *The Measurement of Music and the Cerebral Clock: a new theory*. LEONARDO, 22(1), 1989:83-89.

ROHANI, K., Chen, M.S. & Manry, M.T. *Neural subnet design by direct polynomial mapping*. IEEE Transactions on Neural Nets, 3 (6), 1992:1024-1026.

SEMJEN, A., Schulze, H.-H. & Vorberg, D. *Temporal control in the coordination between repetitive tapping and periodic external stimuli.* Fourth Rhythm Workshop: Rhythm Perception and Production, Bourges, June, France, 1992:73-78.

SOUSA, E.M.G. *Towards an Acoustic Description of Brazilian Portuguese Nasal Vowels*. XIII International Congress of Phonetic Sciences, August 13-19, Stockolm, Sweden, 1995.

TESNIÈRE, L. *Éléments de syntaxe structurale*. Paris: Klincksieck, 1965.

TURVEY, M.T., Schmidt, R.C. & Rosenblum, L. *Clock and motor components in absolute coordination of rhythmic movements*. Status Report on Speech Research SR-101/102, Haskins Labs,1990.

VAN SANTEN, J.P.H. *Assignment of segmental duration in text-to-speech synthesis*. Computer, Speech and Language 8, 1994:95-128.