A MODEL OF SEGMENT (AND PAUSE) DURATION GENERATION FOR BRAZILIAN PORTUGUESE TEXT-TO-SPEECH SYNTHESIS

Plínio A. Barbosa
Laboratório de Fonética Acústica e Psicolingüística Experimental-LAFAPE
Instituto de Estudos da Linguagem
Universidade Estadual de Campinas
CP 6045 - 13081-970 Campinas-SP, Brazil
Tel: +55 19 2397784, FAX: +55 19 2391501, E-mail: plinio@iel.unicamp.br

ABSTRACT

This work presents and evaluates a model of segmental duration generation for Brazilian Portuguese where the notion of macrorhythmic unit is the starting point to drastically simplify duration assignment and to allow pause insertion as an integrated procedure of generation. This model is preferred to random assignment with the same error distribution.

Some aspects of rhythm phonetics and phonology are also discussed that constitute a first step to the understanding of the prosodic component of the language under study.

1. INTRODUCTION

An earlier model for French automatic segmental duration generation was proposed that has represented an important advance in rhythmic structure generators for TTS systems ([1]). That model has allowed to straightforwardly assign segment and pause duration as parts of the same mechanism of generation.

This ability is possible thanks to the notion of a normalized duration , the syllable z-score, as outlined by Campbell ([2]). This notion was extended to IPCG (inter-p-center group) and used within "a neural net plus repartition algorithm" model for generating segment and pause durations. Z-scores' computation for IPCGs including possible silence (as a part of final lengthening in French) has enabled duration assignment and automatic insertion of silent pause. It is important to note that the use of macrorhythmic units has drastically simplified the segmental duration assignment task.

An implicit assumption in our framework is to conceive rhythm production as a by-product of the subject's choice of one among many strategies to perform an underlying metrical structure whose building blocks are syllable-size units called rhythmic programming units (RPU). In French, as the subject speaks, syllables (typically CV sequences) at the first (macro)rhythmic level are performed as IPCGs (typically VC units). In Brazilian Portuguese (henceforth BP), on the other hand, it has been shown ([3]) that at least two macrorhythmic units are necessary to model duration: syllables and IPCGs. In that language, lexical stress can be assigned to the last, penultimate (the most frequent) or antepenultimate syllable (oxyton, paroxytons and proparoxytons, respectively) and the acoustic correlates of stress are often the greater duration of the stressed syllable and the decrease of intensity in the post-stressed ones. Stressed syllables can be enhanced as they are uttered by carrying phrasal accent. Only lexically stressable syllables can bear phrasal accent, where duration and pitch play the major roles in building prominence.

The need for considering two RPUs and the existence of (weak) post-stressed syllables constitute challenges for building an automatic BP segmental duration generator. As in French, the BP duration generator also computes segmental durations in two stages. But in this case, the neural net is conceived in such a way as to sequentially map a phonological, prosodic description for each sentence at the input to the corresponding syllable and IPCG z-scores at the output. The repartition algorithm in the second stage distributes a given duration among IPCG segments bearing phrasal accent and among syllable segments lexically stressed (not phrasally prominent in the sentence). As IPCG consonants include next syllable onsets, the assignment of duration to segments is not a simple task.

2. OVERVIEW OF THE Z-SCORE MODEL

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

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

where Dur_s is 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*. A strong elasticity hypothesis says that all segments in a RPU frame have the same z-score: a single value of *z* per RPU can then be obtained recursively by writing:

$$Dur(RPU) = \sum_s \exp(\mu_s + z.\sigma_s) \quad (2)$$

The log-transformed durations were determined from a 1195-nonsense word *corpus*, containing all BP phonemes (from the São Paulo State dialect). Average and standard-deviation per phone have confirmed (see table 1 in [3]) current knowledge on duration in BP which is in agreement with universal trends.

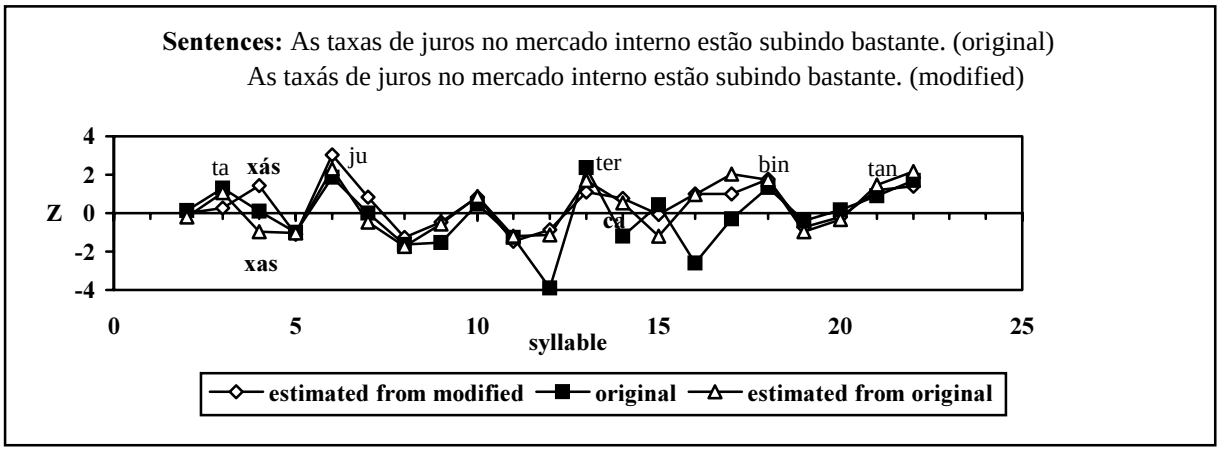

Figure 1: Comparison between the rhythmic patterns for the syllables in the sentence "As taxas de juros no mercado interno estão subindo bastante." (original) and "As taxás de juros no mercado interno estão subindo bastante." (modified). Notice that the accentual peak (learned by the network) on "ta-" (2nd position) migrates to the 3rd position (corresponding to "xás") and that the remaining z-scores for estimated from original and from modified sentences are close to each other.

Another *corpus* was used in order to study rhythmic patterns that emerge from BP sentences. Syllable and IPCG z-scores were computed for 100 sentences read by the same speaker. This *corpus* was manually segmented and carefully labeled by the author. Sentence length varies between one and 84 syllables. Syntactic boundaries were also marked using a set of eight hierarchical labels (see [1] for details).

3. THE RHYTHMIC STRUCTURE GENERATOR

In all 100 rhythmic patterns (represented as durational contours of syllable and IPCG z-scores), 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*. IPCG z-scores delimitate accentual groups (prosodic words) where rhythmic patterns are 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 analyses carried on adjacent segmental z-scores (from formula 1) have confirmed that segment durations are strongly correlated within syllable (at lexical stress) and within IPCG (at phrasal accent) frames.

The observed regularities were used for training a multilayered perceptron. In the network 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 than that of the sequential network used for French (where durations expressed in units of a clock were used instead of z-scores). But original RPU durations are no longer preserved in the BP case. What is preserved is the rhythmic structure as represented by the z-score patterning.

Our model of segmental duration generation was applied to learning and also to test *corpus* subsets. The model was capable to generalize even when lexical stress position was manipulated, as can be observed in figure 1, where the pseudo-word "taxás" has an estimated duration contour coherent with the oxyton pattern.

The example above shows clearly that the network was capable of associating the crucial linguistic information concerning lexical stress to a specific durational patterning.

In the model's next stage, segmental z-scores are computed simply by taking the z-scores previously obtained at the network output and using them in formula 1 for each segment in the sentence. It is important to say that, in BP, two kinds of z-scores must be carefully manipulated. By default, each IPCG z-score is used for computing the durations of the segments that constitute this rhythmic unit. But at lexical stress not enhanced phrasally, the syllable z-score must be used to compute the durations of the segments in the syllable frame. By doing this, the consonants in the syllable onset that are supposed to be part of the previous IPCG would have their durations computed again. To correct this, nucleus and coda segment durations of the previous IPCG are computed by using an averaged IPCG z-score (mean between previous and next IPCG z-scores) and onset segment durations at lexical stress are computed once by using the syllable z-score. Onset segments in the next syllable are not taken into account with this kind of computation. Their duration can be obtained by using the

same z-score of the IPCG sharing the same vowel with the corresponding syllable (a kind of reset value).

With this algorithm, 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.

4. EVALUATION OF THE MODEL

The results presented here show clearly that 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.

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 by analysis-resynthesis with the Hybrid Model ([4]). 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 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 consists of 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 by writing down on a specific sheet.

The results shows 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 still being improved). Three analysis-resynthesis-generated utterances whose durations were obtained by our model can be heard here [sound A0467S01.WAV, A0467S02.WAV, A0467S03.WAV].

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 ([5]) are taken as an approximation, an important amount of the utterances' duration errors for each utterance would be very close or under the threshold. If this assumption is true, it is very hard for the listeners to perceive any difference between the two versions of the original utterance. A certain amount of *don't care* responses reinforces this hypothesis.

5. AUTOMATIC PAUSE INSERTION

A new *corpus,* pronounced by another speaker (40-year-old State of Pernambuco dialect), was recorded to study BP rhythm and, particularly, the problem of silent pause insertion. This *corpus* was also manually labelled and segmented and presents subsets of sentences and nonsense words embedded in carrier sentences uttered at three speech rates (self-chosen normal rate, and metronome-controlled slow and fast rates).

Final lengthening and pause phenomena are challenging in BP. In French, pausing is said to be part of final lengthening. The same can be said for oxytons in BP. But in paroxytons and proparoxytons, prepausal stressed syllables (phrasal accent position, in our terminology) are lengthened and are (often) followed by non-lengthened post-accented ones.The first analyses of the durational contours for this speaker confirmed our previous finding of two macrorhythmic units in BP and seem to indicate that pre-stressed and post-stressed syllable-size units function as a reference clock. At phrasal accent, post-accented syllables are likely to be a kind of filler of a quantized beat interval ([6]). In this position, the IPCG z-score can be computed as follows.

The duration of the entire unit, from current to next vowel onset (including eventual silence) is taken as the starting point. From this duration we extract *n* times the duration of a previously computed reference clock period, where *n* is the number of post-accented IPCGs in the unit. The z-score is computed using (2) with the segments of the accented IPCG.

At the generation stage, if the estimated z-score for that position is greater than a previously assumed value, an amount of "sound" z-score will be computed, as it was done for French. This "sound" z-score is used to compute segment durations for the accented IPCG. Post-accented IPCGs will have their segment durations computed in two steps: z-score computation using a given reference clock period as the left member in (2) and segment duration computation using (1). The silent interval is the amount of duration necessary to complete the next integer number of reference clock periods for the entire unit.

The analyses of the durational contours also show that if duration contours are compared across the three speech rates, some differences between the phonology and the

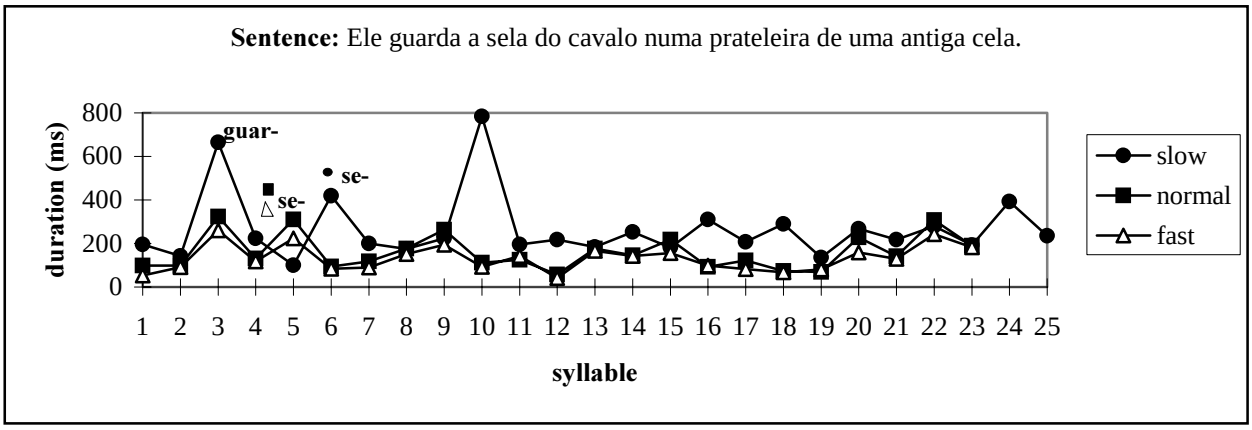

Figure 2: Durational contours for the sentence "Ele guarda a sela do cavalo numa prateleira de uma antiga cela." ("He keeps the horse saddle on a shelf in an ancient cell."). Vertical axis stands for syllable duration in milliseconds and horizontal one, the position of the syllable in the sentence. Please note the evolution of the first four z-scores in the sentence (corresponding to "Ele guarda"). Contour lines are showed only for the sake of visibility.

phonetics of rhythm can be observed that contitute an argument in favor of the initial assumption.

6. CONCLUSION

Clearly stated, this assumption says that rhythm is performed with a continuous phonetic variation subject to physical laws as inertia but can still be understood as the pragmatic-conditioned subjective interpretation of a single metrical representation.

Suppose that the following metrical grid for the first four positions in the sentence in figure 2 is assumed (where *guar-* is phrasally prominent):

```
      x
x     x
x     x
x x   x  x
Ele guarda
```

This grid fits with the slow rate phonetic pattern. As speech rate increases, the need to accelerate the accent realization on *guar-* forces the durational contour to be a *crescendo* towards the strongest position in the grid. This behaviour resembles the entrainment model proposed by Port and colleagues ([7]) if we take the demands of an accentual prominence clock as stronger in a hierarchy than the syllable one (this one would be synchronized to the vowel onset succession). Syllable durations are shown here because they are traditional units, but the pattern shown above is even more *crescendo*-like for IPCG durations and z-scores.

Our rhythm model is part of a project for building a concatenative text-to-speech synthesis system for BP that aims at incorporating well-grounded knowledge of linguistic structure during all research stages (from unit inventory choice [8] to acoustic signal generation). We advocate that high-quality concatenative synthesis is possible once an accurate understanding of the phonetics-to-semantics aspects of a language is assumed.

The careful understanding of phonetic and phonological aspects of rhythm as tentatively presented here is a crucial step in this direction.

ACKNOWLEDGEMENT

This work has been supported by grants (n^{os}. 95/9708-6 and 96/7832-4) from the *State of São Paulo Foundation for the Promotion of Science* (FAPESP).

REFERENCES

[1]Barbosa, P.A. & Bailly, G. (1997) *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. 365-381.

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

[3]Barbosa, P.A. (1996) *At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration.* Proceedings of the 1st ESCA TRW on Speech Production Modeling, Autrans, France, 85-88.

[4]Böeffard, O. & Violaro, F. (1994) *Using a hybrid model in a Text-to-Speech system to enlarge prosodic modifications.* Proceedings of the ICSLP'94, Yokohama, Japan, 727-730.

[5]Fant, G. & Kruckenberg, A. (1996) *On the quantal nature of speech timing.* Proceedings of the ICSLP'96, pp. 2044-2047.

[6]Goedemans, R. & van Heuven, V.J. (1995) *Duration perception in subsyllabic constituents.* Proceedings of the EUROSPEECH'95, Madrid, Spain, 1315-1318.

[7]Port, R., Cummins, F. & Gasser, M. (1996) *A dynamic approach to rhythm in language: Toward a temporal phonology*. In B. Luka and B. Need (eds) CLS-31: Proceedings of the Chicago Linguistics Society, pp. 375-397.

[8]Albano, E.C & Aquino, P.A. (1997) *Linguistic criteria for building and recording units for concatenative speech synthesis in Brazilian Portuguese*. In these Proceedings.