Secondary stress in Brazilian Portuguese: the interplay between production and perception studies

Pablo Arantes and Plinio A. Barbosa

Speech Prosody Studies Group, State University of Campinas, Brazil
pabloarantes@gmail.com

Abstract

This paper reports experiments on speech production showing that secondary stress in Brazilian Portuguese (BP) can be best described as phrase-initial prominence cued by greater duration and pitch accent excursion in initial position. It also reports a perception experiment in which clicks were associated to consecutive V-to-V positions in stress groups. Mean click detection RTs are gradient, but show no influence of initial lengthening. RTs near the phrasally stressed position are shorter and almost 60% of RT variance can be accounted for by produced timing patterns.

1. Introduction

Recent phonological accounts of secondary stress in Brazilian Portuguese (henceforth BP) [1] follow traditional approaches [2] in saying that secondary stress is assigned leftwards to every even syllable counting from the lexically stressed one. Most of the literature on secondary stress in other Romance languages agree on the claim about its binary nature. They have in common the fact that they analyse isolated words, compounds (and in some cases two-word noun-phrases) and rely solely on impressionistic methodologies to base their proposals.

Apart from intuitions based on linguists' introspection, experimental studies have until the moment failed in providing a sound empirical basis for the binary alternation claim in BP (see [3] and others cited in [5]), as well as other Romance languages ([4] for instance). Experimental analysis seem to suggest that some kind of initial prominence cued by greater duration or f_0 excursion is a common feature in this language group. The tendency toward initial prominence is also pointed out in the phonological literature and can be formalized as an iambic reversal rule.

Earlier results [5] of a more comprehensive phonetic study of stress groups containing polysyllabic words with a varying number of prestressed syllables in BP confirmed the initial prominence tendency. Normalized V-to-V duration patterns were compared to those produced by a simulation with a dynamical model of rhythm production [6] and the simulated durational contours mimicked the natural ones quite satisfactorily.

Since rhythm (and other prosodic traits) are traditionally seen as an optimal solution between speakers' and listeners' oposing needs, a more complete account of secondary stress and other rhythm-related phenomena requires a better understanding of the perceptual mechanisms involved in the on-line processing of the patterned signal speakers provide listeners with. Stating the problem in dinamical systems terms it is necessary to uncover how production and perception, each one possessing it own intrinsic dynamics, couple to each other. One way to do this is to determine what parameters listeners are sensitive to when figuring it out what elements in the speech chain stand out as prominent.

The successful interplay between the study of BP timing and its modelling in a dynamical systems framework attained in [5] encouraged us to experimentally investigate the possibility that listeners actively evaluates V-to-V units duration as a cue to detect an upcoming phrasal stress. If this hypothesis comes to be proved true it will be an indication that the the underlying dynamics governing rhythm production and listerners' attention deployment rhythm are similar.

As far as secondary stress goes, the approach we are developing can help answering if the binarity embodied in most phonologists' analysis can be said to play any role in defining the dynamics underlying rhythm perception. The experiment presented in section 3 helps answering this question showing how listeners process the duration patterns elicited in the production study reported in section 2 and in [5].

2. Production Study

Our *corpus* is composed of 17 penultimately-stressed target words, with the number of prestressed syllables ranging from two to five. An independent variable, d_σ, was introduced to control for the possible interplay between phrasal and hypothetical secondary stresses. This variable measures the distance between the syllable bearing lexical stress in the target words and the syllable bearing the main phrasal stress in the NP, in the following carrier sentences: "[A *target* $]_{NP}$ parece menor hoje." ($d_\sigma = 0$) and "[A *target* rude/rural/bicolor $]_{NP}$ parece menor" ($d_\sigma = 2$, 3, 4). A naive male speaker of the southeastern BP variety read ten repetitions of the sentences in a sound-attenuated booth. Duration and f_0 were the main dependent variables.

2.1. Timing Patterns

Segmental duration was grouped in vowel, CV and V-to-V frames (for a reference on the use of V-to-V units for characterization of rhythmic patterns, see [7]) and then normalized by means of z-score transformation to minimize intrinsic duration effects. An *ad-hoc* reference corpus was used for this purpose. No statistical evidence favoring either initial or alternating prominence could be found if individual phone duration was grouped in syllables or if vowel duration was taken alone.

V-to-V duration contours are shown in figures 1 and 2 alongside the first positions in the stress groups for the nine words with four syllables (figure 1) and the six words with five syllables (figure 2). Point markers stand for the different values of d_σ.

Taking consecutive positions in the stress group and different values of the distance parameter (d_σ) as categori-

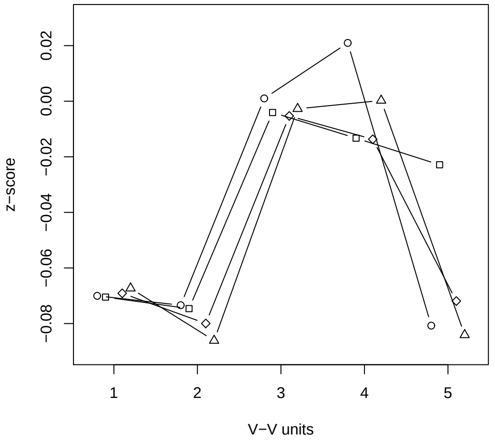

Figure 1: *Normalized duration in stress groups containing four syllable words. Point markers stand for different values of d_σ. ○ is 0, □ is 2, ⋄ is 3 and △ is 4.*

cal factors and mean normalized duration as dependent variable, an ANOVA carried on the group of four-syllable words yields significance only for the position factor ($F(2, 1305) = 456.06,\ p < 10^{-4}$), although positions 1 and 2 are not statistically different as pointed out by a Scheffé post-hoc test. A α level of 5% was adopted.

As for the five-syllable words group, only the position factor yields significance ($F(3, 1144) = 124.3,\ p < 10^{-4}$). Here, positions 1 and 2 are statistically different, according to Scheffé test ($p < 0.003$ for $d_\sigma = 2$ and $p < 0.04$ for $d_\sigma = 4$). These results are evidence for a gradient lengthening of the first V-to-V unit as the stress group gets longer, due to longer target words or the presence of an adjective ($d_\sigma = 4$). No evidence for binary alternations shows up.

2.2. Intonational Patterns

A set of f_0 contour samples from our *corpus* was examined and labeled with the help of three phoneticians. In the agreed transcription, a H^* tone is associated to the first syllable of the target words, irrespective of its length. A complex tone $\text{H}^*{+}\text{L}$ or $\text{L}^*{+}\text{H}$ is associated with the lexically stressed syllable of the target.

Future work on BP intonation will show if pitch accents should be expected in initial position in circumstances other than those involving stress groups starting with polysyllabic words.

2.3. Summing up

It's difficult to see how the results, taken together, can be accounted for by metrical-like representations. Timing and intonational patterns suggest that the initial prominence is related to prosodic phrasing rather than to a hierarchical relation established with the lexically stressed syllable. It seems more appropriate to consider this subordinate prominence as a stress group initial strengthening. Support for this interpretation comes from the fact that initial lengthening can be generated by the rhythm production model [6]. Follow-up studies should investigate how the pattern of prominence(s) along prestressed syllables is affected by (a) target word position within sentence and (b) semantic factors like referential status (i.e., if target word is given or new information).

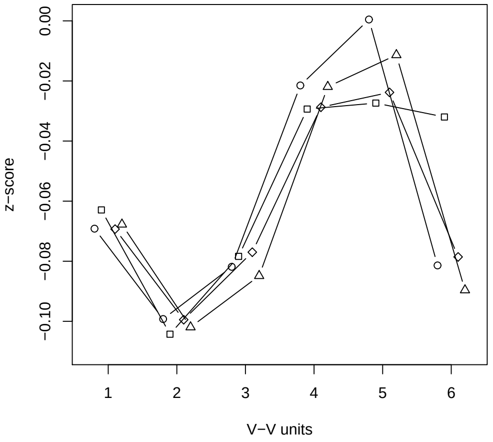

Figure 2: *Normalized duration in stress groups containing five syllable words. Point markers stand for different values of d_σ. ○ is 0, □ is 2, ⋄ is 3 and △ is 4*

3. Perception Study

Duration contours elicited in our production study are highly constrained. The shape of change in duration that culminates in a phrasal stress follows a pattern that can be successfully accounted for by the model described in [6]. Given this fact, the following questions can be raised: **(1)** since prosody is a tradeoff between speakers' and listeners' needs, then listeners' perception of such stimuli are in some way constrained? **(2)** If so, is this pattern related to the one we find in production?

There has been some positive answers to question (1). It has been found that when a word carries sentence stress its perception is somehow facilitated [8] [9]. This finding has been brought to light by having subjects responding to word-initial phoneme targets on stressed and unstressed words and it came out that reaction time (henceforth RT) speeds up when the target phoneme is in a stressed position. The claim is that in every sentence there are points which catch listeners' attention. Therefore, words in those spots are more accurately processed. It has been also demonstrated [10] that in word lists where timing was carefully manipulated so that stressed syllables seemed to occur at periodic points in time, RT to target phonemes were shorter when compared to a situation where some jitter was introduced in the timing of stressed syllables. It seems thus clear that listeners can benefit from rhythmically patterned stimuli.

These experiments cannot state, however, if the effects of timing are local to the rhythmically crucial spots or if these spots are actively exploited all over the stimuli. Studies by experimental psychologists on how people attend to time-changing stimulus [11] suggest the latter option is likely to be true. To put forward the idea that speech perception and production constrain each other is a way to start answering question (2). The hypothesis stated here is that it can be expected that listeners' attention is actively entrained by speakers' activity of producing timing patterns. Since the earlier reported study provided produced timing data and there is a model that can reproduce it, the entrainment hypothesis can begin to be put in test.

3.1. Hypothesis Statement

If the shape of duration is important to the process of perception, there should be a definite relation among changes in duration and changes in the way duration is perceived. In order to verify if this is the case it was investigated how well each consecutive V-to-V unit is attended to and if the changes observed in responding level can somehow be related to changes of duration pattern in the stimuli. Attention was indirectly measured at individual V-to-V unit position in a stress group from the means of reaction time to clicks associated to them.

3.2. Stimuli Preparation

Sixteen sentences were chosen from the production *corpus* that best fitted duration contours in figures 1 and 2. The choice was made so as to pick two phrasing conditions and two target word sizes. The phrasing conditions are illustrated below:

(a) “[A *target*]$_{NP}$ parece menor hoje.” ($d_\sigma = 0$)

(b) “[A *target* bico**lor**]$_{NP}$ parece menor hoje.” ($d_\sigma = 4$)

In condition (a) the target word bears main phrasal stress and in (b) the boldfaced syllable bears main phrasal stress. Eight target words were selected, four with four syllables and four with five syllables. 2.5 kHz pulse-like clicks were added in the original sound files in successive V-to-V positions following the schema below. In (c), target word is “patarata” and in (d) target word is “jaratacaca”. Slashes enclose segments in V-to-V units in orthographic representation. Inside each V-to-V unit, the click was always inserted at the right end of the consonantal segment.

(c) four-syllable words:
/ap/$_1$ /at/$_2$ /ar/$_3$ /**at**/$_4$
/ap/$_1$ /at/$_2$ /ar/$_3$ /at/ a bicol /**orp**/$_4$

(d) five-syllable words:
/aj/$_1$ /ar/$_2$ /at/$_3$ /ac/$_4$ /**ac**/$_5$
/aj/$_1$ /ar/$_2$ /at/$_3$ /ac/$_4$ /ac/ a bicol /**orp**/$_5$

For each word in each phrasing condition four — as in (c) — or five — like in (d) — clicked sound files were generated. A total of 72 test sentences were generated applying this procedure. An additional contextualizing sentence was added prior to each test sentence. Another 68 pairs of sentences were recorded by the same speaker of the production *corpus* to be used as fillers. A click was randomly placed always in the first sentence on filler items.

3.3. Stimulus Presentation

Items consisting of a pair of sentences were presented to the subjects who were instructed to hear them and answer yes or no to a content question following stimuli presentation. Subjects were also warned that at some point during their listening of the pair of sentences a click would appear and they were instructed to press a joystick button as fast as they could after the click. The answer to the yes or no question was also recorded by pressing a joystick button.

DMDX software was used to present the audio files and record RT to click monitoring. Instructions were presented in written form to each subject and a training section was run in the presence of the experimenter so that any doubts could be solved. Items were randomized prior to each round and blocked in three parts with breaks between blocks. Sound files were presented over closed headphones in a quiet room. Eighteen college students voluntarily participated in the experiment and signed a consent term. Subjects took about 20 minutes to do the entire test.

3.4. Statistical Procedures

Raw RT data was log-transformed and then z-score normalization was applied according to the following procedure: mean RT ($\bar{X}_j$) and standard deviation (s_j) was computed for each subject j and then each i RT sample of j (x_{ij}) was transformed into z_{ij} by means of equation 1.

$$z_{ij} = \frac{x_{ij} - \bar{X}_j}{s_j} \qquad (1)$$

There are two reasons to apply such operations. First, RT is taken here as a measure of attention deployment and we are interested in differences due to treatment and not in absolute values of RT. Besides, log-transformation and z-score normalization help fitting highly right-skewed distributions such as RT raw data into the normal distribution.

Separate two-way ANOVAs were run in two groups: (A) the test sentences with four-syllable target words and (B) the group with five-syllable target words. Position in the stress group and phrasing condition were independent variables and normalized RT was the dependent variable. A α level of 5% was fixed for statistics carried out in the experiment.

3.5. Results

Figures 3 and 4 show mean normalized RT for each V-to-V position in groups (A) and (B) respectively.

For group (A), factors Position ($F(3,509) = 10.688,\ p < 10^{-5}$) and Phrasing ($F(1509) = 4.0343,\ p < 0.05$) reached significance level but not the interaction. *Post-hoc* testing with Scheffé shows that position 4 (phrasal stress bearer) has statistically lower mean than positions 1 ($p < 10^{-5}$) and 2 ($p < 0.02$). Likewise, position 3 it is significantly lower than position 1 ($p =< 0.03$) yet marginally higher that position 4 ($p < 0.09$).

As for group (B), only Position factor yielded significance ($F(4,632) = 5.0132,\ p < 10^{-5}$). *Post-hoc* Multiple comparisons with Scheffé indicate position 5 (phrasal stress bearer) and 2 are significantly different ($p < 0.05$). Besides, position 4 differs significantly from position 2 ($p < 0.02$) and marginally from position 1 ($p < 0.06$).

It is worthy noting that position 2 in figure 4 has a slightly higher mean RT compared to position 1, yet this difference is non-significant. The same comparison reached significance when duration means were compared (cf. section 2.1). It is possible, though, that in longer stress groups a larger RT difference between first and second positions shows up.

These results amount to classical findings that during sentence processing significant reduction in RT is shown in rhythmically crucial points, which can be seen as a positive answer to question (1) stated above. In addition, these results, specially those concerning group (A), bring new evidence that perception facilitation is a gradient process not confined to rhythmically salient spots. The closer a V-to-V unit is to stress group boundary (as represented by duration maxima in figures 1 and 2), the

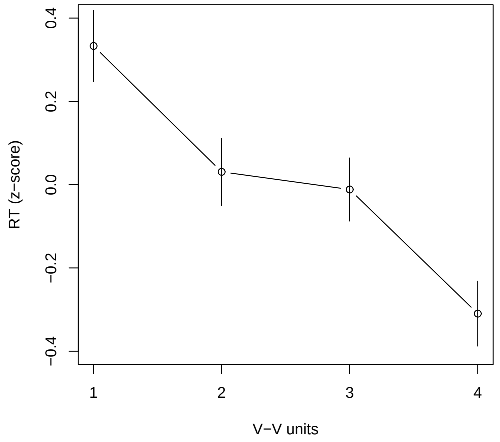

Figure 3: *Normalized RT for click detection along positions in stress groups of four V-to-V units. Whiskers indicate standard error.*

faster it is attended to. It means that perception facilitation is somehow a function of time, what can be seen as a preliminary evidence that question (2) can also have a positive answer.

The issue to be settled now is whether time is to be interpreted just as order (V-to-V position in stress group) or if the perceptual entrainment is related to actual produced timing. In order to tackle this problem, duration and RT data were correlated.

Mean V-to-V duration of the four V-to-V units of the eight sentences in group (A) were correlated to the corresponding normalized RT means. The best correlation was achieved through non-linear estimation using a polynomial function like $y = a + bx + cx^2$ yielding $r = -0.764$ $(p < 0.009)$. Duration data explain 58% of RT variance in this case.

As for group (B), means of position 1 to 3 were pooled over and so were means of position 4 and 5 as suggested by Scheffé homogeneous groups test. RT means were pooled over the same way. Likewise group (A), the best correlation was the one we got through non-linear estimation using the same function. In this case $r = -0.77$ $(p < 0.002)$ and a proportion of 59% of RT variance can be accounted for by duration data.

4. Discussion

Correlation results seem to represent preliminary evidence that speech perception and production patterns can in fact be closely related, since almost 60% of RT variance in the experiment can be traced back to duration scaffolding in production. Future experiments are to show if correlates of intonation play any role in predicting perception. Besides that, they should also investigate how semantic information helps listeners predict when and where speakers will place stress along a sentence.

5. Acknowledgements

This work was supported by FAPESP grants 03/11619-0 and 05/02525-7. We thank Luciana Lucente and Sandra Madureira for their help with intonational labelling.

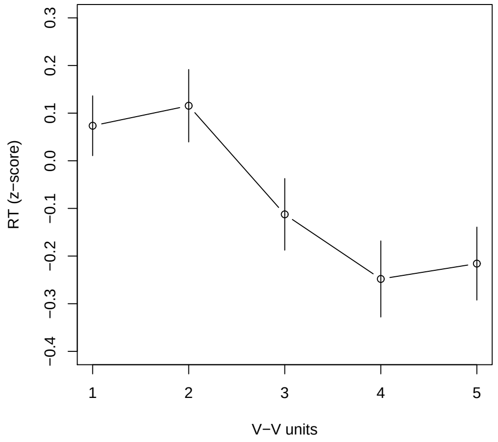

Figure 4: *Normalized RT for click detection along positions in stress groups of five V-to-V units. Whiskers indicate standard error.*

6. References

[1] Collischonn, G., 1994. Acento secundário em português brasileiro. *Letras de Hoje,* 29, 43–53.

[2] Said Ali, M., 1908. *Difficuldades da Lingua Portugueza: Estudos e Observações.* Rio de Janeiro: Laemmert.

[3] Moraes, J. A., 2003. Secondary stress in Brazilian Portuguese: perceptual and acoustical evidence. In *Proceedings of the 15th ICPhS.* Barcelona, Spain, 2063–2066.

[4] Prieto, P.; van Santen, J., 1999. Secondary stress in Spanish: some experimental evidence. In *Aspects of Romance Linguistics.* C. Parodi et al. (eds.). Washington: GUP, 337-356.

[5] Barbosa, P. A.; Arantes, P.; Silveira, L. S., 2004. Unifying stress shift and secondary stress phenomena with a dynamical systems rhythm rule. In *Proceedings Speech Prosody 2004,* Nara, Japan, 49–52.

[6] Barbosa, P. A., 2002. Explaning cross-linguistic rhytmic variability via a coupled-oscillator model of rhythmic production. In *Proceedings of Speech Prosody 2002.* Aix-en-Provence, France, 163-166.

[7] Barbosa, P. A.; Bailly, G. 1994. Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, 15: 127-137.

[8] Cutler, A.; Foss, D. N., 1977. The role of sentence stress in sentence processing. *Language and Speech*, 20(1), 1–10.

[9] Martin, J. G., 1972. Rhythmic (hierarquical) versus serial structure in speech and other behavior. *Psychological Review*, 79(6), 487–509.

[10] Quené, H.; Port, R. F., 2005. Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica,* 62(1), 1–13.

[11] Large, E. W.; Jones, M. R., 1999. The Dynamics of Attending: How People Track Time-Varying Events. *Psychological Review*, 106(1), 119–159.