ISCA Archive
http://www.isca-speech.org/archive

Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

Production–perception entrainment in speech rhythm

Pablo Arantes and Plinio A. Barbosa

Department of Linguistics, State University of Campinas, Brazil
pabloarantes@gmail.com, pabarbosa.unicampbr@gmail.com

Abstract

The article reports favorable initial results to the hypothesis that rhythm perception can be seen as a listener–speaker entrainment process. The data comes from an experiment in which subjects had to detect a click in test sentences. Each sentence contained one click that was associated to one of the syllables of two consecutive stress groups defined by duration criteria. Reaction time (RT) to click detection is assumed to reflect the degree of listener-speaker entrainment: faster detection meaning stronger entrainment. Results show that the closer to the phrasally stressed syllable the click is the faster the RT is. The crucial result concerning our working hypothesis, though, is that RT slows down after the stress group boundary, resuming its decrease trend afterwards. Multiple linear regression analysis performed on different acoustical parameters of the test sentences shows that duration and F_0 explain around 50% of RT variance. We interpreted these results as a positive preliminary corroboration of the entrainment hypothesis by showing that boundaries in the spoken utterances seem to trigger a reset in entrainment activity and that duration seems to be the main acoustical feature driving listeners' behavior.

Index Terms: speech rhythm, rhythm modeling, rhythm processing

1. Introduction

We are interested in unifying the approach to speech rhythm production and perception within the framework of dynamical models. More specifically, we aim at bridging the gap between existing models of the production and perception sides of the speech rhythm phenomenon.

McAuley [1] advances the idea that it is possible to model auditory rhythmic pattern *perception* employing the adaptive-oscillators formalism to account for the fact that people get entrained by the temporal structure of many different events in the environment. The model does not address, however, the specificity of spoken language timing structure.

Barbosa [2], on the other hand, puts forward a speech timing *production* model also based on the adaptive-oscillators framework. In his model, complex timing patterns similar to the ones found in natural speech are generated as the outcome of coupling between oscillators representing two hierarchical levels of organization: phrase stress beats and the regular flow of syllable-sized units.

These examples demonstrate that the same mechanisms, coupling and entrainment, can be used to model different aspects of the speech rhythm phenomenon, suggesting that a unified approach to speech rhythm is possible. Barbosa [3], for instance, acknowledges that the underlying idea behind McAuley's model [1] could in principle be used to explain *speech* rhythm perception as long as it is extended to incorporate Barbosa's idea that two mutually influencing oscillators rather than just one are needed to account for the specificities of speech timing.

Saltzman and colleagues [4] have proposed an oscillator-based account of prosodic influence on articulation that is compatible with the unifying approach we are seeking. The authors, in agreement with Barbosa, acknowledge the convenience of resorting to entrainment mechanisms such as those employed by McAuley's model to account for both production and perception of complex temporal structures within an oscillator-based framework. Also as Barbosa, Saltzman and colleagues *assume* the production–perception entrainment process without explicitly modeling it, basing their assumption on models such as McAuley's and Large's [5]. Since those models does not primarily address *speech* timing perception, it may be unsafe to assume they will be able to model speech data without revisions. In this scenario, it becomes important to clarify whether existing entrainment models are in fact comparable to the actual behavior of listeners.

An early attempt to obtain such evidence was carried out by Krivokapić [6], who tried to find a relation between perceived boundary strength (PBS) estimated by listeners and articulatory events at the vicinity of prosodic boundaries. According to Saltzman and colleagues' framework these articulatory events show the influence of temporal modulating gestures. Krivokapić's results are not very conclusive in either direction, probably because the PBS paradigm, an offline, highly metalinguistic task, requires listeners to consciously attend to a very precise point in the sentence, what may not be comparable to the kind on online processing of whole and complex patterns that is characteristic of real spoken language. In designing the experiment reported here we tried to avoid metalinguistic tasks.

2. Hypothesis statement

Barbosa's model [2] implements period-coupling between the phrase stress and syllabic oscillators, such that ongoing pulses of the first oscillator cause the period of the second to lengthen. The model also specifies syllabic oscillator period resetting after the phrase stress pulse. Altogether, it means that: (i) rhythm production is organized around stress groups (defined as the interval between two phrase stresses) and (ii) syllabic oscillator period behavior can function as an important signal of the location of a boundary.

The listener–speaker entrainment hypothesis postulates, based on (i), that rhythm perception is also organized around stress groups, and, based on (ii), that listeners are able to track changes in syllabic oscillator period to help them locate stress group boundaries. Given those two postulates, the hypothesis makes the testable prediction that there should be a resetting behavior similar to the one the production model implements at the perception level.

This prediction can be tested by measuring listener's readi-

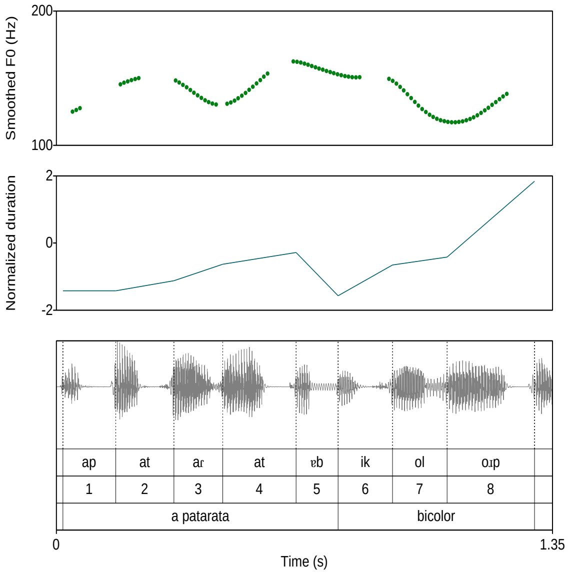

Figure 1: Smoothed F_0 and normalized duration contours for test sentence “A patarata bicolor parece menor.” Bottom panel shows waveform and syllable-sized unit boundaries. Decreased duration of 5$^{\text{th}}$ syllable-sized unit signals stress group boundary.

ness in tracking syllable-sized units along a stress group as an indication of how entrained listeners are to the speaker’s speech. In the experiment reported here the tracking readiness will be measured by the listener’s ability to detect clicks associated to successive syllables along two consecutive stress groups.

The main hypothesis of the experiment is that click detection latency will be increasingly short within a stress group, although, there will be a local rise in detection times around the boundary. This is what Martin [7] calls the “scallop effect”.

3. Stimuli

Test sentences are of the form (vertical bars indicate stress group boundary):

(1) A patarata |bicolor |parece menor.
Det. Noun Adj.
‘The two-colored fool seems smaller’

Test nouns are four four-syllable and four five-syllable penultimate stress words. In each group two words have [a] as their syllable nuclei in all pre-stressed syllables and other 2 have [o]. Sentences with four-syllable nouns have eight target positions (cf. Figure 1) and those with five-syllable nouns have nine. Each original sentence generated eight or nine test sentences with one click in one of the test positions. Clicks were inserted at successive intervocalic intervals midpoints.

4. Experimental design

Target positions along the stress groups are the independent variable: eight target positions for the sentences with four-syllable nouns (cf. Figure 1) and nine target positions for those with five-syllable nouns. Reaction time (RT) to click detection is the dependent variable.

To ensure subjects are in fact listening to the sentences and not just waiting for the clicks they were asked to answer a yes/no question about each sentence meaning upon hearing it.

5. Experimental procedure

Subjects were exposed to 67 experimental items (one missing target position due to a ˈCV.V sequence in one test word) and 73 filler items (filler items also had one click each) with an inter-item interval of 800 ms presented in four blocks with self-administered pause in between. Stimuli were presented using the DMDX [8] software. Reaction time was collected with a Logitech G5 mouse with USB polling rate of 500 Hz to minimize delay errors. The subjects were 42 Brazilian Portuguese native speakers subjects and they took around 20 minutes to complete the experiment.

6. Statistical analysis

Box-Cox transformation [9] was applied to raw RT data to reduce non-normality and assimetry. A semi-automatic procedure was used to find the optimal λ for the whole of the sample data:

$$y_i^{(\lambda=0.8)} = \frac{y_i^\lambda - 1}{\lambda} \tag{1}$$

Transformed RTs were then normalized by subject’s mean and standard deviation because we are interested in differences due to treatment and not in absolute values of RT:

$$z_{ij} = \frac{y_{ij} - \bar{Y}_j}{s_j} \tag{2}$$

Separate one-way Anovas were run in the eight and nine target positions groups. POSITION is the independent variable and transformed-normalized RT is the dependent variable. Multiple comparisons had their p-values corrected using the Bonferroni method. A 5% α level was fixed for all tests. The R statistical environment [11] was used for all statistical analysis.

7. Results

7.1. Eight target positions condition

Figure 2 shows mean normalized RT for each of the eight positions along the two target stress groups. There’s a significant effect of POSITION on RT means $[F(7, 1170) = 17.830,\ p < 0.001]$. Pairwise comparisons show that for positions 2–3 and 4–8 all intra-group comparisons are non-significant and all inter-group are significant ($p < 0.01$), except for position 6 for which all intra- and inter-group comparisons are non-significant.

That RT contour over the consecutive target positions indicate a gradient decrease in RT as the positions go further into the sentence. There seems to be evidence, though, that the stress group boundary between positions 5 and 6 causes a local increase in RT.

7.2. Nine target positions condition

Figure 3 shows mean normalized RT for each of the nine positions along the two target stress groups. There’s a significant effect of POSITION on RT means $[F(8, 1270) = 11.3,\ p < 0.001]$. Pairwise comparisons show that there are no significant

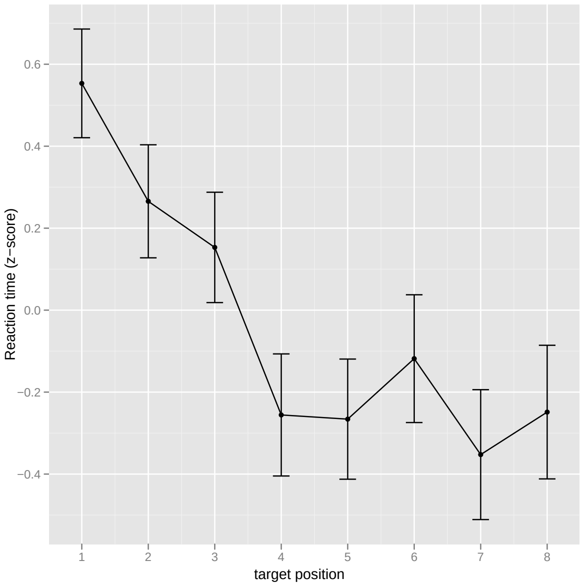

Figure 2: Mean normalized RT for each position along two stress groups. Stress group boundary between positions 5 and 6. Whiskers indicate 95% confidence intervals around the mean.

differences between positions 5-7-8-9, but comparison between positions 6–7 and 6-9 show significance ($p < 0.02$).

As with the other condition, RT also decreases rightward into the sentence up to the stress group boundary when there's a local RT increase. The difference is that in this case the "scallop" is one position earlier in comparison to the other condition.

7.3. Regression analysis

In order to assess how strongly the different acoustic features of the test sentences influence RT data, multiple linear regression analysis were run. The acoustic parameters investigated were:

- duration: raw, normalized and smoothed with a 5-point moving average function
- F_0 (measured in the voiced part of each target position): median value and rise
- spectral emphasis (following methodology outlined at [10])

For each target position the mean RT (all subjects pooled) was correlated with four different values: the acoustic parameter value for the previous, the current and the following target position and also the discrete derivative of the acoustic parameter (the difference between the current and previous value of the parameter).

The best multiple regression analysis was obtained with smoothed duration of the previous position and median F_0 derivative as independent variables ($R^2 = 0.5$, $p < 0.001$). Figures 4 and 5 show the relationship between the acoustic parameters duration and F_0 and RT along with best-fit lines. They reveal that RTs for a given target position tend to be slower when the duration of the previous target position is small. That makes sense when the RT contours are compared with the duration contours: the "scallop" on RT observed at position 6 (Figure 2) takes place after the duration drop observed between positions 4 and 5 (Figure 1). As for F_0, RT for a given position tends to be faster when the present target position median F_0 is lower than the previous position a negative difference between the current and the previous position median F_0 value. This relation is not so obvious when comparing Figures 1 and 2.

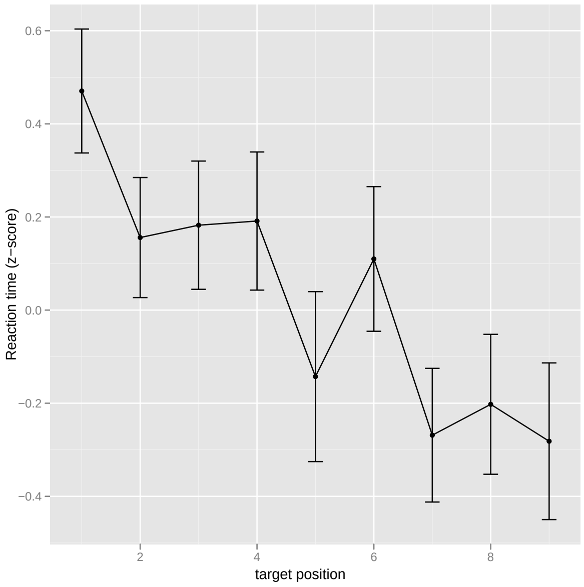

Figure 3: Mean normalized RT for each position along two stress groups. Stress group boundary between positions 6 and 7. Whiskers indicate 95% confidence intervals around the mean.

8. Discussion

There's a great deal of variability in the results and the differences between position means are not as clear as someone might hope for, probably due to individual differences in performance, to the fact that RT data is inherently noisy and also due to uncontrolled aspects of the stimuli. The variability makes it unclear if the perceptual entrainment reset takes place after stressed syllable (which seems to be the case of the nine target positions condition) or after the end of the morphological/phonological word (which is the case in the other condition).

Nevertheless, by pointing out that RT to clicks associated to target positions in the vicinity of stress group boundaries tend to be detected slower than in positions within the stress groups the experiment results seem to provide initial evidence for the listener-speaker entrainment hypothesis by showing that (i) listeners track syllable-sized duration units when listening to speech and (ii) as in speech rhythm production, prosodic boundaries have an active role organizing listener's experience of rhythm.

In order to further test the hypothesis there are variations of the present experiment that can be done. One possibility is to manipulate the prosodic boundary strength to test how it correlates with different levels of post-boundary RT increase either by constructing different test sentences were the test stress groups are separated by a stronger syntactic break or by doing resynthesis of the test sentences. Other possibility is to replicate the experiment with low-pass filtered versions of test sentences in order to assess how much of the listener's behavior can be

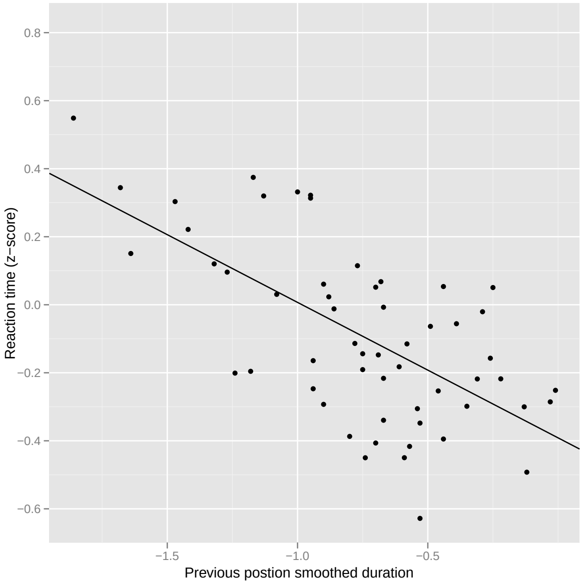

Figure 4: Reaction time as a function of smoothed duration of previous target position. The straight line represents the linear regression (intercept = -0.4, slope = -0.4).

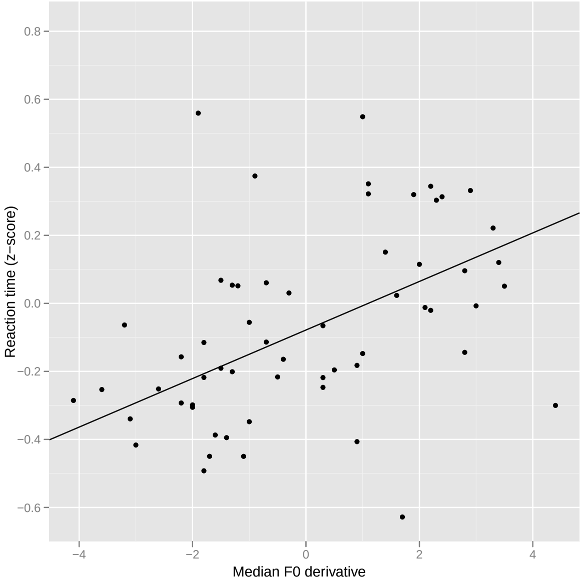

Figure 5: Reaction time as a function of median F_0 derivative. The straight line represents the linear regression (intercept = -0.08, slope = 0.07).

attributed to lexico-grammatical knowledge.

9. Acknowledgements

Work supported by a FAPESP PhD scholarship (06/50166-9). Special thanks to our speaker and subjects.

10. References

[1] J. D. McAuley. (1995). *Perception of time as phase: Toward an adaptive-oscillator model of rhythmic pattern processing.* PhD thesis, Indiana University, Bloomington.

[2] P. A. Barbosa. (2007). From syntax to acoustic duration: a dynamical model of speech rhythm production. *Speech Communication*, 49(1–2): 75–96.

[3] P. A. Barbosa. (2006). *Incursões em torno do ritmo da fala*. Campinas: Pontes.

[4] E. Saltzman et al. (2008) A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proceeding of the Speech Prosody Conference 2008*, 175–184.

[5] E. W. Large and M. R. Jones (1999) The dynamics of attending: How people track time-varying events. *Psychological Review*, 106, 119–159.

[6] J. Krivokapić. (2007) The planning, production, and perception of prosodic structure. Unpublished PhD thesis, Univesity of Southern California.

[7] J. G. Martin. (1986). Aspects of rhythmic structure is speech perception. In J. Evans and M. Clynes (Eds.), *Rhythm in psychological, linguistic, and musical processes*. Springfield, IL: Charles C. Thomas.

[8] K. I. Forster and J. C. Forster. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers,* 35(1): 116–124.

[9] G. E. P. Box and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B 26 (2): 211252.

[10] A. M. C. Sluijter and V. J. van Heuven. (1996). Spectral balance as an acoustic correlate of linguistic stress. *J. Acoust. Soc. Am.*, 100, 2471–2485.

[11] R Development Core Team. (2007). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.