INTERSPEECH
2009 BRIGHTON

Cross-variety Rhythm Typology in Portuguese

Plínio A. Barbosa[1], *M. Céu Viana*[2], *Isabel Trancoso*[3]

[1]Speech Prosody Studies Group/Dep. of Linguistics/Inst.Est. Ling., Univ. of Campinas, Brazil
[2]Center of Linguistics of the University of Lisbon, Portugal
[3]INESC-ID, Lisbon, Portugal

pabarbosa.unicampbr@gmail.com, mcv.fono@gmail.com, Isabel.Trancoso@inesc-id.pt

Abstract

This paper aims at proposing a measure of speech rhythm based on the inference of the coupling strength between the syllable oscillator and the stress group oscillator of an underlying coupled oscillators model. This coupling is inferred from the linear regression between the stress group duration and the number of syllables within the group, as well as from the multiple linear regression between the same parameters and an estimate of phrase stress prominence. This technique is applied to compare the rhythmic differences between European and Brazilian Portuguese in two speaking styles and three speakers per variety. Compared with a syllable-sized normalised PVI, the findings suggest that the coupling strength captures better the perceptual effects of the speakers' renditions. Furthermore, it shows that stress group duration is much better predicted by adding phrase stress prominence to the regression.

Index Terms: rhythmic variability, rhythm typology, coupled-oscillators

1. Introduction

The issue of finding a way to characterise speech rhythm from acoustic measures is one of the most persistent enterprises in speech prosody research. This traces back to a period when American Spanish and American English were put in two different sides as regards rhythm typology: respectively syllable-timed and stress-timed. As the alleged absolute regularity of syllables or stressed syllables was never found in production, the current mainstream of research in prosody moved its efforts to search for evidence of rhythm classes on a gradient acoustic space [14, 10, 12, 1, 5].

In broad terms, two schools of rhythm research can be identified. For one of them, which follows a modelling framework, the complexity of durational patterns found in natural languages is a consequence of the coupling between hierarchically organised oscillators [13, 1, 7], usually a stress group oscillator and a syllable oscillator. For the other school, following a descriptive framework, the proportion and variability of non-successive vocalic and consonantal intervals [14] (%V and ΔC Indexes) or the degree of discrepancy between successive vocalic and consonantal intervals [11] or syllabic intervals [12, 8] (Pairwise Variability Indexes: PVI) would describe crucial cross-linguistic rhythmic differences.

The work in [11, 10] showed the advantage of the PVI indexes over Ramus' proposals, by controlling for speech rate and by taking into account local measures of acoustic intervals. However, only indexes based on vocalic intervals (nPVI-V) seem to be both resistant to influences of speech rate [15] and more robust to changes of speaker and speaking style [10]. Recently, measures based on syllabic intervals were used for describing rhythmic differences across varieties of English [12, 8]. Despite the authors' comments, indexes based on these intervals seem to be as effective as vocalic intervals' indexes in distinguishing speech rhythm types (see tables 1 and 2 in [12]).

The main goal of the present work is to compare the success of syllable-sized PVI indexes with measures of coupling strength between syllable- and stress group oscillators to measure rhythmic differences due to language variety and speaking style. The varieties compared are European (henceforth EP) and Brazilian Portuguese (henceforth BP), and the speaking styles, read text and story telling. From a phonetic and phonological point of view, both varieties may be considered as rhythmically mixed as they have a relatively simple syllable structure, like most syllable-timed languages, but they also exhibit a high frequency of vowel reduction, a property most generally associated with stress-timed ones. Vowel reduction is more extreme in EP than in BP, as unstressed reduced vowels are shorter, more centralized and more often deleted. Crucially, vowel deletion rates for the three EP speakers used in the present work range from 25 to 31 % over all underlying vowels, whereas the BP subjects exhibit a rate of vowel deletion from 15 to 24 %. EP is expected, then, to be closer to stress-timed languages than BP. This is suggested by [9] in an experiment replicating [14] for these two varieties: in a %V x ΔC space, EP clusters with British English and Dutch, whereas BP is at the opposite side of that space. This extreme separation of BP from EP rhythms is challenged, however, by the results of discrimination experiments, in which the utterances were low-pass filtered. Then, two sets of stimuli were presented, one with the intonational contour preserved and the other, with a flattened contour. EP native speakers could only discriminate between the two varieties in the former condition, but did so in both conditions for pairs of different languages.

Before presenting the corpora and the findings, a brief overview of some aspects of coupled oscillators theory and a technique for computing the coupling strength are presented.

2. Explaining the complexity of patterns of duration with coupled-oscillators models

Hierarchical models of speech rhythm production, such as [3, 13], posit from start the necessary coupling between two or more interacting oscillators, such as between the syllable and the stress-group oscillators. This coupling allows to explain both universal and language-specific properties of syllable-sized patterns of duration by means of general principles of production applicable to all languages, which are language-specifically parameterised by a coupling strength variable.

Among these models, that of [13] allows a simple way to infer the coupling strength parameter value. The Averaged Phase Difference technique is applied to infer the coupling strength value, provided that two conditions be satisfied: (1) that the coupling forces in both directions are symmetrical and differing only in sign and in the coupling strength of the stress group oscillator onto the syllable oscillator, and (2) that the consequences of the coupling for both oscillators derive solely from these bidirectional forces, and from the number of cycles of the faster oscillator within the cycle of the slower oscillator. The authors showed that the coupling strength c between the two oscillators is equal to the ratio between the intercept and the inclination of the linear regression computed between two variables: I and n in equation 1. I is the duration of the stress group; ω_{sg} is the frequency of the stress group oscillator; ω_{σ}, the frequency of the syllable oscillator; n is the number of syllables within the stress group; and $H(\Phi(n))$ the coupling function.

$$I = \frac{1}{\omega_{sg} + H(\Phi(n))} = \frac{c}{c.\omega sg + \omega_{\sigma}} + n.\frac{1}{c.\omega sg + \omega_{\sigma}} \quad (1)$$

This proposal represents a paradigm change (see also [7]) in speech rhythm research, because it allows to restate data on isochrony in relative terms: the higher the coupling strength, the more stress-timed a language is, and vice-versa. There is no need to refer to any kind of absolute isochrony. Indeed, provided that both regression coefficients are significant, a value of 1 for $c = intercept/inclination$ stands for an even influence between both oscillators. On the other hand, if $0 < c < 1$, the syllable oscillator dominates the stress group oscillator (syllable timing), and if $c > 1$, the stress group oscillator dominates the syllable oscillator (stress timing). Distinct languages or varieties, as well as speakers and speaking styles would differ in degree of coupling, but not in nature of the underlying phenomenon. Rhythmic classes can be found a posteriori by statistical analysis of coupling strength variability.

O'Dell and Nieminen's technique was applied in this work and compared with the values of a syllable-sized PVI to assess the rhythmic structure of six speakers in two Portuguese varieties. After that, a change in perspective is proposed, which introduces the need for including estimates of phrase stress prominence in the regression analysis.

3. Methodology

3.1. Corpora

The corpora consist of parallel productions of six subjects in both EP and BP. Two native female and one native male speakers for each variety read a text on the origin of the pastries *pastéis de Belém* (reading style, RE). After the reading, the six subjects told what the text was about (story telling style, ST). Each native speaker read the text written in his/her own written variety. All speakers aged 30 to 45 years, and were full or student researchers on speech science and technology.

As the stories told by some speakers was much shorter than the reading material, excerpts containing a little more than 300 words were chosen for analysis in the twelve productions (six speakers times two speaking styles).

3.2. Delimiting stress groups automatically

The vowel onsets for all readings were marked semi-automatically in Praat [6]: automatic VO detection by the Beatextractor script [3], followed by manual correction. Each interval delimited by two consecutive VOs defines the so-called VV unit. More than $3,450$ VV units were then segmented and tagged with a broad phonetic transcription.

Stress groups were delimited by automatically detecting phrase stress boundaries within and across connected utterances. The sequence of phrase stress positions was automatically tracked by serially applying two techniques for normalising the VV durations. The first one was a $z - score$ transform: $z = \frac{dur - \sum_i \mu_i}{\sqrt{\sum_i var_i}}$, where dur is the VV duration in ms, and the pair (μ_i, var_i) are the reference mean and variance in ms of the phones within the corresponding VV unit. These references are found in [2, p. 489] for BP. For EP, a reference table was created from the analysis of a corpus of read speech. This transform was followed by a 5-point moving average filtering: $z^i_{smoothed} = \frac{5.z^i + 3.z^{i-1} + 3.z^{i+1} + 1.z^{i-2} + 1.z^{i+2}}{13}$. In both BP and EP phrase stress is placed at the right edge of the duration-related stress group. The normalisation technique, and the detection of duration-related phrase stress boundaries from the detection of $z^i_{smoothed}$ maxima were implemented by a Praat script (SGdetector). The computation of both the stress group duration and the number of VV units in the stress group was made by the SGdetector script. The counting of the number of phonological syllables in the stress group was done manually.

Since the procedure is entirely based on duration maxima, the right boundary not necessarily coincides with a lexically stressed unit. Sometimes a post-stressed lengthened VV unit signals the end of the stress group. Silent pauses were included in the VV units that precede them. In doing so, high values of z-scores were obtained from VV units containing silent pauses, signalling a strong prosodic boundary. Offglides were included in the VV unit containing the vowel leftwards. Onglides formed a vocalic unit with the vowel rightwards.

In both BP and EP, as signalled by [4] for Romance languages, stress groups usually contain on average 7 syllables, with groups with up to 15 phonological syllables or more.

3.3. Estimating the coupling strength between oscillators

In order to estimate the coupling strength between the underlying syllable and stress group oscillators, it is necessary to compute the linear regression between the duration of the stress groups (I), and the number of syllable-sized units (n) of the corresponding stress groups. As explained above, the coupling strength is the ratio between the intercept coefficient (a) and the inclination coefficient (b) in the regression equation $I = a + b.n$. Since variation is proper to language, speech rhythm variation is expected to occur across speaking styles, individuals, languages, varieties, among others. All these kinds of variation are explored in this paper.

Departing from O'Dell and Nieminen's proposal, the relevance of the level of prominence of each phrase stress for explaining the variance of the stress group duration is taken into account in this work. Besides the number of syllable-sized units, the prominence level of the stress group is taken as the second independent variable for predicting the stress group duration. The prominence level was taken as the smoothed z of each phrase stress (see previous section). Due to the techniques applied, this value is not a subpart of the stress group duration.

3.4. Estimating the syllable-sized nPVI

A normalised Pairwise Variability Index ($nPVI_{VV}$) was computed for each analysed excerpt, according to the formula 2, as proposed by [11]. The index is the sum, for all stress groups of the excerpt, of the discrepancy between two consecutive values of raw duration d_i and d_{i+1} within a stress group of length m, normalised by the averaged duration of the units compared. Instead of vowels or consonants, the VV unit is used, because it is taken as the minimum rhythmic unit. Other works, such as [12, 8], also used a syllable-sized unit for computing PVI (in those cases, the phonological syllable).

$$nPVI_{VV} = \sum_{SG_k} \frac{1}{m-1} . \sum_{i=1}^{m-1} \frac{|d_i - d_{i+1}|}{mean(d_i, d_{i+1})} \qquad (2)$$

If an excerpt is stress timed, for which a higher discrepancy of consecutive durations is alleged, the PVI is high. Due to the alleged differences between BP and EP rhythms, the former is expected to have lower values of $nPVI_{VV}$ then the latter.

4. Results

As the number of VV units in the stress group did not turn out to produce significant values for some intercept coefficients, only the linear regressions taking the number n of phonological syllables (PU) as an independent variable are shown here. Table 1 shows the linear regression equations, according to language variety (BP/EP), speaker and sex (LLF stands for speaker LL, female, etc), as well as speaking style (RE or ST). The coupling strength c is the ratio between the intercept and the inclination coefficients of the respective equation. All correlation coefficients and (consequently) the inclination coefficient are highly significant ($p < 10^{-4}$). The significancy of the intercept coefficient is indicated between parentheses. In case of non significant values for the intercept coefficient, c is considered undefined (u), except in cases of marginal significance. Quantitative speech rate (sr) is given in PU/s. All correlation coefficients are higher than 0.38.

Table 1: I x n regression equations, coupling strength c and speech rate (sr). I stands for stress group duration and n, for the number of phonological syllables. See text for details.

lang	reg. equation	c	sr
BP-LLF-RE	$I = 310 + 158.n$ ($p < 0.02$)	2.0	5.1
BP-LLF-ST	$I = 170 + 217.n$ (ns)	u	4.2
BP-AGF-RE	$I = 47 + 199.n$ (ns)	u	4.9
BP-AGF-ST	$I = 578 + 166.n$ ($p < 0.002$)	3.5	4.1
BP-FAM-RE	$I = 277 + 150.n$ ($p < 0.09$)	1.8	5.4
BP-FAM-ST	$I = 274 + 195.n$ ($p < 0.2$)	1.4	4.3
EP-SVF-RE	$I = 309 + 160.n$ ($p < 0.1$)	1.9	5.1
EP-SVF-ST	$I = 1020 + 117.n$ ($p < 10^{-3}$)	8.7	4.7
EP-AJM-RE	$I = 174 + 141.n$ ($p < 0.1$)	1.2	6.2
EP-AJM-ST	$I = 785 + 94.n$ ($p < 10^{-4}$)	8.4	5.7
EP-ITF-RE	$I = 156 + 169.n$ (ns)	u	5.3
EP-ITF-ST	$I = 846 + 115.n$ ($p < 10^{-6}$)	1.9	4.8

The results of coupling strength from Tab. 1 are plotted against speech rate in Fig. 1. Undefined values of coupling strength were considered as zero. A striking difference between the story telling and reading styles in EP may be observed: coupling strength is much higher in the former style than in the latter in this variety. In BP, only the female speaker AG has a visibly higher coupling strength when telling the story in comparison with reading it. When reading, both EP and BP seem to exhibit a similar amount of stress timing. When telling a story, a spontaneous situation, EP seems to be much more stress-timed than BP (when listening to excerpts of the material in the two styles, it is also the impression we have. Listen to the audio files BP-AGF-RE.wav, BP-AGF-ST.wav, EP-SVF-RE.wav, EP-SVF-ST.wav). A cultural difference seems to be related to two form of story telling: EP speakers know *pastéis de Belém* quite well, which motivated them to sum up the story in a few words and tell their experience related to the *pastéis*. BP speakers, on the other hand, retold the story in detail. The memory load of EP speakers is thus much higher than the memory load of BP speakers in this condition.

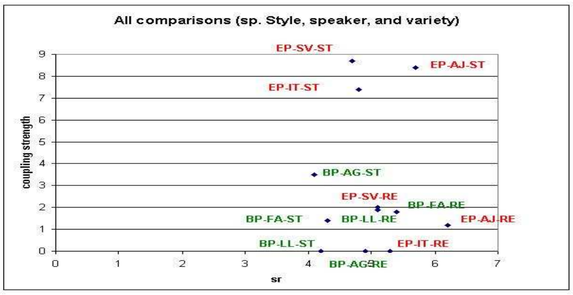

Figure 1: *Relationship between speech rate (in PU units) and coupling strength for BP and EP. Speaking style (RE and ST), and speaker are indicated. Labels are the same of Tab. 1*

A regression including the estimate of the phrase stress magnitude dominating the corresponding stress group reveals, indeed, a better prediction for I, as can be seen in Tab. 2. All correlation coefficients are between 0.79 and 0.97 (that is, 62 % to 94 % of the variance of the stress group duration is explained from the number of phonological syllables combined with the estimated phrase stress prominence).

Table 2: I x n, p regression equations, and ratio a/b for EP and BP. p stands for phrase stress magnitude. Speech rate is the same of Tab. 1. See Tab. 1 and text for details.

lang	reg. equation	a/b
BP-LLF-RE	$I = 215 + 126.n + 63.p$ ($p < 0.02$)	1.7
BP-LLF-ST	$I = -10 + 182.n + 45.p$ (ns)	u
BP-AGF-RE	$I = 71 + 153.n + 62.p$ (ns)	u
BP-AGF-ST	$I = 373 + 138.n + 41.p$ ($p < 0.005$)	2.7
BP-FAM-RE	$I = 197 + 125.n + 63.p$ ($p < 0.05$)	1.6
BP-FAM-ST	$I = 237 + 143.n + 45.p$ ($p < 0.07$)	1.7
EP-SVF-RE	$I = 128 + 131.n + 156.p$ ($p < 0.06$)	2.0
EP-SVF-ST	$I = 441 + 124.n + 78.p$ ($p < 10^{-2}$)	1.0
EP-AJM-RE	$I = 103 + 126.n + 145.p$ ($p < 0.1$)	0.8
EP-AJM-ST	$I = 319 + 101.n + 131.p$ ($p < 10^{-2}$)	3.2
EP-ITF-RE	$I = 79 + 135.n + 161.p$ (ns)	u
EP-ITF-ST	$I = 346 + 124.n + 109.p$ ($p < 10^{-2}$)	2.8

These findings signal the importance of taking the magnitude of prosodic boundaries into account in speech rhythm research. The analysis reveals that both predictors n and p contribute independently to predict I (cross-variable R^2 is inferior to 0.003 for all cases). If the a/b ratio is used in the equations in Tab. 2 as a measure of coupling strength, the general picture is exactly the same of the one shown in Fig. 1. If this result demonstrates the robustness of the simple linear regression in revealing essential aspects of rhythmic typology, it also shows that the coupling strength can be inferred from an equation that better reflects the variance of stress group durations.

Fig. 2 plots nPVI$_{VV}$ against speech rate for all comparisons. Observe that this index appears to normalize for speech rate and that the two speaking styles are not clearly separated. Note also that, although EP speakers may exhibit slightly higher nPVI$_{VV}$ values, BP and EP renditions overlap at their edges. Although such an overlap is somewhat unexpected, it is in agreement with the fact that EP native speakers cannot reliably discriminate between the two varieties, as shown in [9]. Most noticeable, however, is the fact that except for one BP speaker in one speaking style, the distribution of values along the nPVI$_{VV}$ axis in Fig. 2 for both varieties coincide with the that observed for British English by [8], also using a syllable-sized PVI.

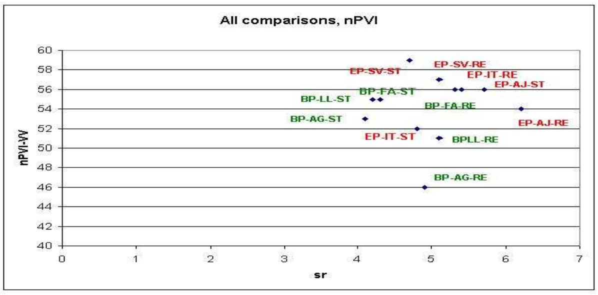

Figure 2: *Relationship between speech rate (in PU units) and nPVI for BP and EP. Speaking style (RE and ST), and speaker are indicated.*

5. Discussion

The findings reveal that both the normalized PVI and the coupling strength computed here for VV intervals predict contiguous positions in a rhythmic space for the two varieties of Portuguese analysed here.

The coupling strength seems, though, more tuned with the subjects' rhythmic performance as it reflects differences in speaking styles. A way of how to formally measure this tuning still needs to be presented, however. Both the number of phonological syllables and the phrase stress magnitude explain more than 60 % of stress group variability in the twelve renditions. Besides including in the analysis linguistic crucial prosodic units such as number of syllables, and phrase stress prominence, other advantages of the framework outlined here are: (1) there is no need to separate vowel intervals from consonant intervals; (2) the consequences of phrase stress magnitude to stress group duration are implemented in a direct way, revealing the importance of the structuring aspect of rhythm production; (3) the tendency to stress timing can be identified with high levels of coupling strength; (4) the tendency to syllable timing can be identified with low levels of coupling strength; (5) both universal and language-specific aspects of speech rhythm can be easily identified in this framework: all languages share the two kinds of oscillators and hence they are prone to exhibit both tendencies towards stress and syllable timing, although different patterns of syllable-sized durations are found due to differences in coupling.

6. Acknowledgments

The first author acknowledges a grant from CNPq (300371/2008-0). This work was also supported by FCT project PTDC/PLP/72404/2006. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III". The authors thank Juva Batella for adapting the text to BP, and all the speakers for their time.

7. References

[1] Barbosa, P. A., "Explaining Cross-Linguistic Rhythmic Variability via a Coupled-Oscillator Model of Rhythm Production", Proc. Speech Prosody 2002 Conf. [CD], Aix-en-Provence, 163–166, 2002.

[2] Barbosa, P. A., Incursões em torno do ritmo da fala, Campinas: RG/Fapesp, 2006.

[3] Barbosa, P. A., "From syntax to acoustic duration: a dynamical model of speech rhythm production", Speech Communication, 49:725–742, 2007.

[4] Beckman, M. E., "Evidence for speech rhythms across languages" in Tohkura, Y. et al. [Eds], Speech perception, Production and linguistic structure, 457–463, IOS Press, 1992.

[5] Bertinetto, P. M., Bertini, C., "Towards a unified predictive model of Natural Language Rhythm", Quaderni del Laboratorio di Linguistica della SNS, 7, 2007/08.

[6] Boersma, P., Weenink, D., "Praat: doing phonetics by computer" (Version 5.0.35) [Computer program], Online: http://www.praat.org, accessed in 2008.

[7] Cummins, F., Port, R., "Rhythmic constraints on stress timing in English", J. Phon., 26:145–171, 1998.

[8] Deterding, D., "Letter to the Editor. The measurement of rhythm: a comparison of Singapore and British English", J. Phon., 29:217–230, 2001.

[9] Frota, S., Vigário, M., Martins, F., "Language discrimination and rhythm class: Evidence from Portuguese", Proc. Speech Prosody 2002 Conf. [CD], Aix-en-Provence, 315-318, 2002.

[10] Grabe, E., "Variation adds to prosodic typology", Proc. Speech Prosody 2002 Conf. [CD], Aix-en-Provence, 2002.

[11] Low, E. L., Grabe, E., Nolan, F., "Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English", Language and Speech, 43:377–401, 2000.

[12] Mok, P. P. K., Dellwo, V., "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English", Proc. Speech Prosody 2008, Campinas, 423-426, 2008.

[13] O'Dell, M., Nieminen, T., "Coupled Oscillator Model of Speech Rhythm", Proc. XIVth ICPhS, San Francisco, 1075-1078, 1999.

[14] Ramus, F., Nespor, M., Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition, 73:265–292, 1999.

[15] Russo, M., Barry, W., "Isochrony reconsidered. Objectifying relations between Rhythm Measures and Speech Tempo", Proc. Speech Prosody 2008, Campinas, 419–422, 2008.