RESEARCH REPORT

USING A COUPLED-OSCILLATOR MODEL OF SPEECH RHYTHM TO ESTIMATE **RHYTHMIC VARIABILITY IN TWO BRAZILIAN PORTUGUESE** VARIETIES (CE AND SP)

Pablo ARANTES
Universidade Federal de São Carlos (UFSCar)

Ronaldo Mangueira LIMA JÚNIOR
Universidade Federal do Ceará (UFC)

OPEN ACCESS

EDITORS
- Miguel Oliveira, Jr. (UFAL)
- René Almeida (UFS)

REVIEWERS
- Rosane Silveira (UFSC)
- Plinio Barbosa (UNICAMP)

ABOUT THE AUTHORS
- Pablo Arantes
  Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.
- Ronaldo Mangueira Lima Júnior
  Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.

DATES
- Received: 08/20/2021
- Accepted: 08/30/2021
- Published: 09/04/2021

HOW TO CITE
ARANTES, Pablo; LIMA JÚNIOR, Ronaldo Mangueira (2021). Using a Coupled-Oscillator Model of Speech Rhythm to Estimate Rhythmic Variability in Two Brazilian Portuguese Varieties (CE and SP). *Cadernos de Linguística*, v. 2, n. 4, e577.

ABSTRACT

This paper presents preliminary results of a semi-automatic methodology to extract three parameters of a dynamic model of speech rhythm. The model attempts to analyze the production of rhythm as a system of coupled oscillators which represent syllabicity and phrase stress as levels of temporal organization. The estimated parameters are the syllabic oscillator entrainment rate (alpha), the syllabic oscillator decay rate (beta), and the coupling strength between the oscillators (w0). The methodology involves finding the <alpha, beta, w0> combination that minimizes the distance between natural duration contours (restored from normalized and smoothed raw duration) and simulated contours generated using several combinations of the parameters. The distance between natural and model-generated contours was measured in two ways by comparing: (1) the plain/overt syllable-to-syllbale duration in the natural contour with that of the model-generated contour and (2) the relative change along both contours. We applied this methodology to read

DOI 10.25189/2675-4916.2021.V2.N4.ID577 ISSN: 2675-4916 V. 2, N. 4, 2021

speech produced by five speakers of the state of Ceará (CE) and eight speakers of the state of São Paulo (SP). Mean w0 and alpha values are compatible with the view that Brazilian Portuguese is a mixed-rhythm language. Results from two Bayesian hierarchical regression models do not suggest a difference between SP and CE speakers, but indicate a difference between the two methods, with the relative change method generating lower alpha values and higher w0 values.

RESUMO

O artigo apresenta resultados preliminares de uma metodologia semiautomática para a extração de três parâmetros de um modelo dinâmico do ritmo da fala. O modelo propõe analisar a produção do ritmo como um sistema de dois osciladores acoplados que representam a silabicidade e acentuação como níveis de organização temporal. Os parâmetros estimados são a taxa de indução do oscilador silábico pelo oscilador acentual (alfa), a taxa de decaimento do oscilador silábico (beta), e a força de acoplamento entre os dois osciladores (w0). A metodologia consiste em encontrar a combinação <alfa, beta, w0> que minimiza a distância entre contornos de duração natural de enunciados e contornos simulados gerados usando as combinações de parâmetros. A distância entre contornos naturais (duração restaurada a partir do contorno normalizado e suavizado) e os gerados pelo modelo foi medida de duas maneiras: (1) a duração propriamente dita de sílaba a sílaba do contorno natural e do simulado são comparadas (2) a mudança relativa da duração ao longo de ambos os contornos é comparada. Aplicamos a metodologia a enunciados lidos produzidos por cinco falantes do Ceará e oito falantes de São Paulo. Os valores médios de w0 e alfa são compatíveis com a análise segundo a qual o português brasileiro é uma língua de ritmo misto. Os resultados de dois modelos de regressão hierárquicos bayesianos não sugerem uma diferença entre os falantes de CE e SP, mas indicam uma diferença entre os dois métodos, com o método da mudança relativa gerando valores menores de alfa e valores maiores de w0.

KEYWORDS

Prosody; Speech Rhythm; Brazilian Portuguese.

PALAVRAS-CHAVE

Prosódia; Ritmo Linguístico; Português Brasileiro.

INTRODUCTION

The goal of this study is to test a methodology we are developing to study rhythmic variability across speakers and languages based on the concepts underlying Barbosa's dynamic model of speech rhythm (BARBOSA 2006; 2007). In this initial exploration of this methodology, we use it to compare rhythmic variability in two Brazilian Portuguese regional varieties, the one spoken in the Northeastern state of Ceará and the one spoken in the Southeastern state of São Paulo.

We base our work on the version of the model presented in detail in Barbosa (2006, 2007). Here we omit the mathematical details of the implementation for the sake of brevity and focus on a qualitative presentation of the key ideas underlying it. The model assumes the existence of two abstract oscillators. The first is a syllabic oscillator, which stands for the sequence of syllable-sized units (or VV units) that make up the speech chain. The second oscillator is the phrase stress oscillator and represents the sequence of beats or phrase stresses that occur along a given utterance. The phrase stress oscillator entrains the syllabic oscillator, such that its period gets progressively lengthened as it approaches the phrase stress beat location in the utterance. Points of maxima on the entrained oscillator cycle correspond to the onsets of vocalic gestures in speech.

Four model variables are relevant to our discussion on rhythmic variability: T_0, the syllabic oscillator uncoupled period, which can be understood as the underlying speech rate; α, the syllabic oscillator entrainment rate, which modulates how fast the syllabic oscillator period changes in response to coupling; β, the decay or period reset rate, which modulates how fast the syllabic oscillator tries to return to the uncoupled period after the phrasal stress beat; and w_0, the relative coupling strength, which modulates the degree of synchronization between the two oscillators.

Barbosa (2006, 2007) makes a connection between his model and the so-called rhythm typology or the classification of a particular language as syllable-timed or stress-timed. This typology has a long tradition in linguistic research – see Bertinetto (1989) for a review. At first, Pike (1945) treated the issue as a matter of a flexible tendency towards one or the other rhythmic type, but later on Abercrombie (1967) reframed the subject as a dichotomy. Even though the empirical research has been unsupportive of the dichotomy view on the matter (BERTINETTO, 1989; DAUER, 1983), the idea has persisted, augmented by the inclusion of moraic rhythm as a third type (LADEFOGED, 2001). Barbosa (2002, 2004, 2006, 2007) suggests that the dynamical nature of his model allows the syllable-timed and stress-timed typology to be reinterpreted as a continuum rather than a discrete dichotomy. He argues, based on theoretical and empirical grounds, that by varying w_0 from low (near zero) to high (near 1) values, it is possible to simulate duration patterns which go from syllable-timed to stress-timed. The author hypothesizes that w_0

values “vary more extremely from language to language, than from speaker to speaker within a same linguistic community” (BARBOSA, 2007, p. 733), but points out that the claim of higher crosslinguistic variation compared to intralinguistic variation should be verified experimentally. He also suggests (Barbosa 2007, p. 734) that w_0 should vary less than α and β within a language community and that α and β may vary as a function of both speakers and speaking styles.

The main goal of the present study is to take a step forward in defining a semi-automatic methodology that will make it possible to estimate α, β and w_0 given a set of audio samples, making it easier to investigate some of Barbosa’s claims about the ability of the dynamic model of speech rhythm to explain both rhythmic typology as well as claims about where particular languages lie within this typology. In this paper we present an outline of the methodology as it stands now, and apply it to speech samples from two Brazilian Portuguese varieties. The results reported here will help to further develop our methodology and enrich the description of BP rhythmic variability. In the spirit of open science and reproducibility, we made all data and script files used in this study available at https://osf.io/w82ru/.

1. MATERIALS AND METHODS

1.1. SPEECH MATERIAL

Two sets of oral production data were analyzed, one from Ceará (CE) and the other from São Paulo (SP). The recordings analyzed are not identical because the data come from different research projects; however, both are readings of texts at a normal speech rate, which allow for the comparison at hand.

The speech material of the CE speakers consisted of five recordings of a 225-word-long text, which was a translation to Portuguese of the diagnostic reading passage from Celce-Murcia et al. (2010). The mean duration of recordings was 87.6 seconds, ranging from 80 to 95 seconds. Speakers (four males and one female) were aged between 18 and 20, all born and raised in the metropolitan area of the state of Ceará.

The speech material of the SP speakers consisted of eight recordings of a 144-word-long text, the passage “A Menina do Narizinho Arrebitado”[1], written by Monteiro Lobato. The mean duration of recordings was 33.5 seconds, ranging from 21 to 54 seconds. Speakers (five males and three females) were aged between 18 to 30, all born and raised in the state of São Paulo.

[1] Both passages can be found in the Appendix.

1.2. PHONETIC ANALYSIS

Speech samples were segmented into vowel-to-vowel (VV) units[2] using a Praat script (ARANTES, 2021) and manually adjusted if necessary. Segments present in VV units were labelled in a *TextGrid* file using the SAMPA-PB convention[3]. Figure 1 shows a *TextGrid* file with the first few seconds of an audio sample segmented in VV units according to the procedure described earlier in the section. Segmentation was used to identify stress groups in speech samples. Stress groups were identified using the 3-step procedure described in Barbosa (2006, 2007) and implemented as a Praat script[4]. First, the raw duration of each VV unit is normalized using a *z*-score transform; in the second step, the normalized duration contour is smoothed using a 5-point moving average; lastly, peaks in the normalized, smoothed contours are identified. Peaks on the contour are considered occurrences of phrasal stresses and two consecutive phrasal stresses define a stress group.

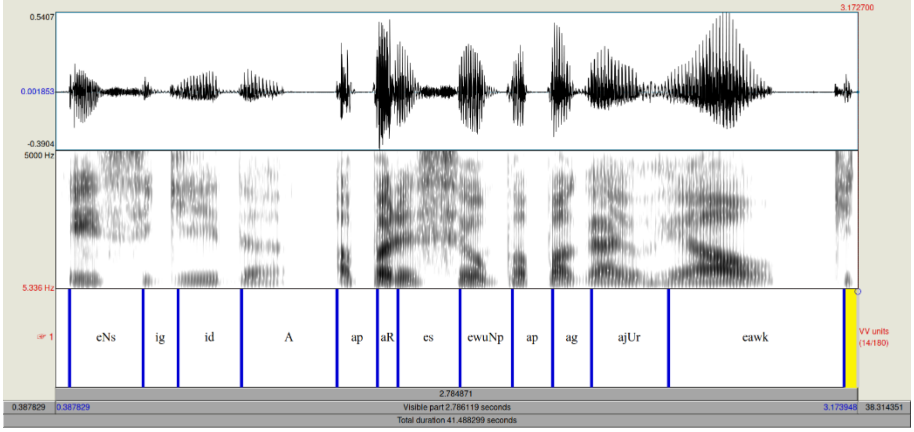

Figure 1. Part of an audio sample from the SP variety (speaker SP1) segmented into VV units following the SAMPA-PB notation. The content shown in the sample is "Em seguida, apareceu um papagaio real" (*Then a royal parrot appeared*).

1.3. MODEL PARAMETER EXTRACTION

The semi-automatic methodology to estimate the parameters involves generating thousands of simulated contours with varying α, β and w_0 values, which are later compared to the natural contours to arrive at the <α, β, w_0> combination that best fits the natural

[2] A VV unit is a syllable-sized segment delimited by two consecutive vowel onsets in running speech. A body of literature suggests VV units are better than the phonological syllable to reveal the rhythmic structure of speech (BARBOSA, P. A., 2006, 2007; PETTORINO *et al.*, 2013).

[3] SAMPA-PB is a convention inspired by SAMPA to phonetically annotate Brazilian Portuguese using ASCII characters. See https://github.com/parantes/sampa-pb for more information.

[4] The script can be found at https://github.com/parantes/duration_suite. It is a rewrite of the SGDetector script originally coded by Barbosa (2006). The reader is referred to Barbosa (2006) for further information on the rationale for each step of the procedure.

contour. Syllable-sized durations generated by the model are expressed in time units. The normalized smoothed natural contours are expressed in *z*-score units. In order to make a comparison between the two types of contours possible, both must be in the same scale. To achieve this, we adopted the procedure described in Barbosa (2004, p. 42-43) to restore the normalized and smoothed duration of each VV unit in the contour to an abstract duration expressed in real time units by reversing the *z*-score transform. The restored abstract duration of a given VV unit (dur_{abs}) is obtained by the use of the formula in (1), where z_{sm} stands for the normalized and smoothed duration of the VV unit, σ_r is the reference standard deviation value and μ_r is the reference mean value.

$$dur_{abs} = z_{sm} \cdot \sigma_r + \mu_r \qquad (1)$$

Barbosa suggests the use of the abstract VV unit "aC" as a reference in the restoration operation. The rationale is that "a" is the most frequent vowel in BP and "C" is a voiceless stop consonant (in BP, those are [p t k]), the most frequent consonantal class in BP. The value of μ_r is obtained by summing the reference mean value of [a] and the mean of the reference mean values of [p t k], giving 212 ms as a result. The value of σ_r is obtained by taking the square root of the sum of the reference variance values of the four segments, which gives 56 ms as a result. The mean and standard deviation values used as reference are given in Barbosa (2006).

In order to generate simulated contours using the coupled-oscillator model, it is necessary to specify, in addition to the α, β and w_0 parameters, an estimate of the uncoupled syllabic oscillator period (T_0), the number of stress groups, the number of VV units in each stress group and the magnitude of each phrase stress. The information on the number of stress groups and their size in VV units is provided by the output of the *duration-suite* script mentioned in the previous section. T_0 was estimated by taking the median duration of VV units in stress groups with more than four VV units, excluding the first two and the last VV units. The first two VV units are excluded because the model stipulates that the syllabic oscillator period reset is most active at the start of the stress group. The last VV unit is excluded because it is the most affected by the influence of phrase stress. The magnitude of phrase stresses was kept constant at a value of 1. The rationale for this decision is that the formulation of the coupled-oscillator model we use here does not incorporate information on higher linguistic information that may affect boundary strength and especially the presence and duration of silent pauses at the boundary[5].

[5] Barbosa (2006, 2007) presents a probabilistic approach to incorporate the effect of the syntax-prosody interaction on phrase stress magnitude, although it does not deal with silent pause insertion.

When generating simulated contours, the information on the number of stress groups and their size (in VV units) was kept constant and values for the α, β and w_0 parameters are systematically varied: α and β vary from 0.05 to 1.5 in steps of 0.05, and w_0 varies from 0.05 to 1 in steps of 0.025. In doing so, a total of 35,100 simulated contours are generated based on the stress group structure of each speech sample, one for each <α, β, w_0> combination. In order to generate the simulated contours, an R implementation of the the coupled-oscillator model was used[6].

1.4. CONTOUR COMPARISON AND RANKING

In order to determine which <α, β, w_0> combination generates the simulated contour that best matches a natural contour, the natural contour is compared to all simulated contours. The distance between contours was measured using two methods:

- *Plain duration*: the restored abstract duration (expressed in real time units) of each VV unit along the natural contour is directly compared to the corresponding unit in the model-generated contour.
- *Relative change in duration*: instead of duration itself, the series of unit-to-unit relative change in duration is compared. Relative change is computed by dividing the backward difference (i.e., the duration of the i^{th} position in the contour minus the duration of the (i - 1) position) by the i^{th} position duration.

The *plain duration* method (Figure 2) is the most straightforward way to compare the natural and simulated contours. The rationale for creating the *relative change* method (Figure 3) comes from Barbosa (2004). In that paper, the author outlines a procedure based on rearrangements of his model's mathematical formulas that allows estimating w_0 by measuring relative change in duration contours[7].

[6] The code can be accessed at https://github.com/parantes/rhythm.

[7] In the procedure described in Barbosa (2004), the value for α is fixed by the user, but the value may change for different speaking rates if the data to which the procedure is being applied systematically vary the speaking rate.

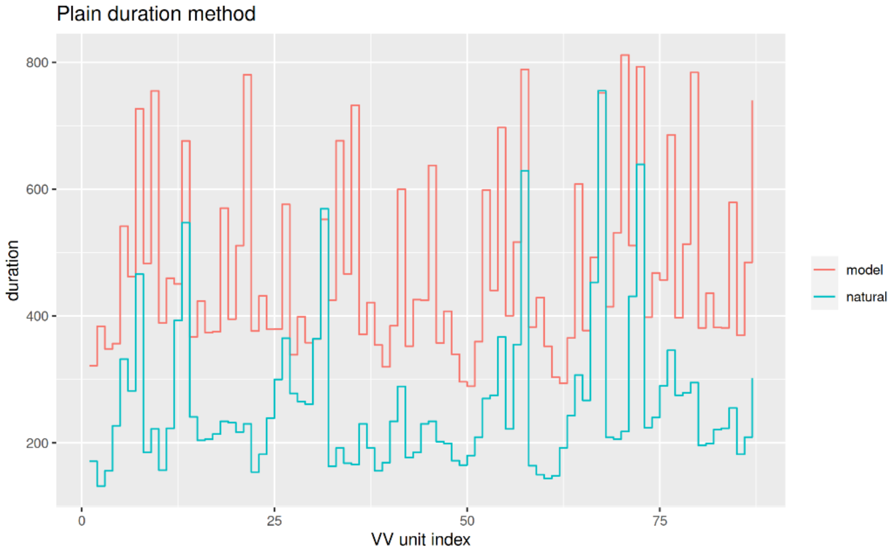

Figure 2. Example of natural and simulated contours being compared using the plain duration method. The simulated contour is one of the best ranked in the comparison with the natural contour. α, β, w_0 and T_0 values are 1, 0.45, 0.5 and 224 ms (Speaker SP2).

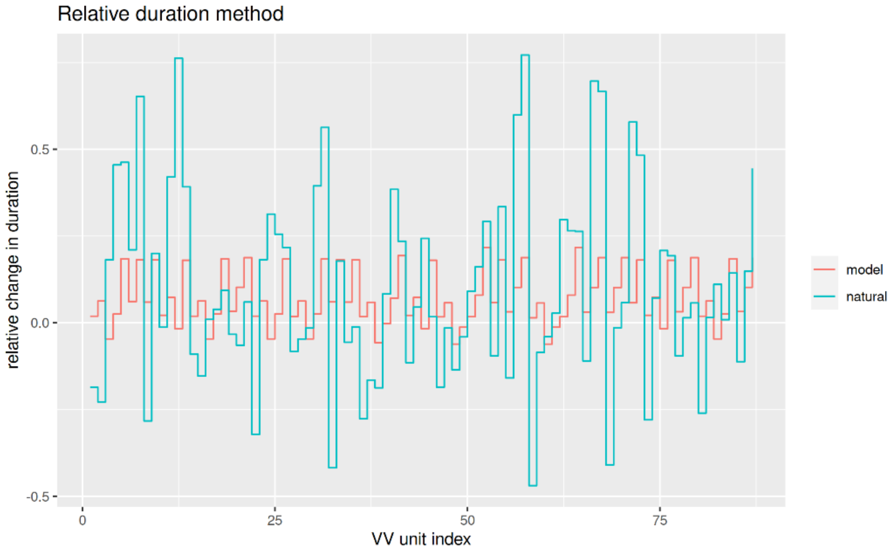

Figure 3. Example of natural and simulated contours being compared using the relative change in duration method. The simulated contour is one of the best ranked in the comparison with the natural contour. α, β, w_0 and T_0 values are 0.3, 0.85, 0.6 and 224 ms (Speaker SP2).

Also following Barbosa (2004), only stress groups with more than four VV units were included in the comparisons; and, within the stress groups, the first and the last VV units were excluded. The rationale for this is the same outlined in the previous section for the estimation of T_0.

Mean Absolute Error (MAE) was used as the measure of distance between natural and simulated contours[8]. Error was measured as the mean of the absolute difference in each VV unit between the corresponding values in the natural and simulated contours. Simulated contours that are impossible or unlikely in natural speech were excluded from the comparison, namely contours that have negative values. Contours that have a minimum that is smaller than the minimum VV duration found in the natural contour by more than 10% or a maximum that exceeds the maximum VV duration in the natural contour by more than 10% were also excluded.

In addition to the error measure, the ratio between the duration range of natural and simulated contours was used as a second index to rank a contour comparison. Range ratios close to 1 ensure that the simulated contour spans a range of VV unit durations that is closer to the one found in the natural contour.

In order to determine the parameter combination that yielded the simulated contour with the best fit to the natural one, the list of all valid simulated contours was sorted in an ascending order by the range ratio index. A list with the 50 candidates with ratios closest to 1 was generated and then sorted in descending order of error measure. The median of the values of α, β and w_0 of the first 20 candidates in this list are considered the values that generate the simulated contour that best fits the natural contour.

2. RESULTS

Since β, out of the three parameters, is the one with the least impact on rhythmic typology, and for the sake of brevity, this analysis consisted of two response variables: α and w_0; and two predictor variables: method (*plain duration* and *relative change*), and dialect (CE and SP). Table 1 presents the mean and standard deviation of α and w_0 for each combination of predictor variables.

[8] We also tested using Dynamic Time Warping (DTW) and Root-Mean-Square Error (RMSE) as error measures. All three measures performed well, and we decided to use MAE since it generated less variability.

	CE		SP		ALL	
	α	w_0	α	w_0	α	w_0
plain duration	$\bar{x}$ = 0.93 s = 0.132	$\bar{x}$ = 0.438 s = 0.11	$\bar{x}$ = 1.022 s = 0.128	$\bar{x}$ = 0.516 s = 0.125	$\bar{x}$ = 0.987 s = 0.133	$\bar{x}$ = 0.486 s = 0.121
relative change	$\bar{x}$ = 0.855 s = 0.072	$\bar{x}$ = 0.652 s = 0.1	$\bar{x}$ = 0.825 s = 0.155	$\bar{x}$ = 0.702 s = 0.112	$\bar{x}$ = 0.837 s = 0.126	$\bar{x}$ = 0.683 s = 0.106

Table 1. Mean ($\bar{x}$) and standard deviation (s) of α and w_0 for each dialect and for all speakers together, under the *plain duration* and *relative change* methods.

Both methods yielded w_0 values that are consistent with the hypothesis of Brazilian Portuguese having a mixed rhythm, with the *relative change* method generating higher values. As can be seen from table 1 and figure 4, the method that generated higher values for α (*plain duration*) had lower values for w_0, and vice-versa: the method that generated lower α values (*relative change*) had higher w_0 values. The grand mean for both dialects pooled together for α was 0.987 under *plain duration* and 0.837 under *relative change*; and for w_0 was 0.486 under *plain duration* and 0.683 under *relative change.*

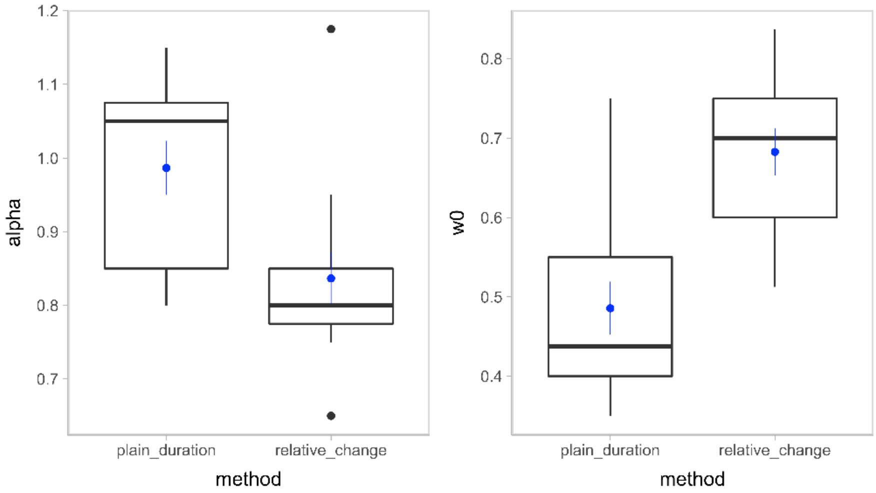

Figure 4. Boxplots of α and w_0 values of CE and SP dialects pooled together for the *plain duration* and *relative change* methods. Blue dots and blue lines are the means and standard errors, respectively.

Considering the dialects separately, as shown in Figure 5, SP speakers had higher w_0 values in both methods, and also had higher α values in *plain duration*. The only parameter that was higher for CE speakers was their α values under the *relative change* method.

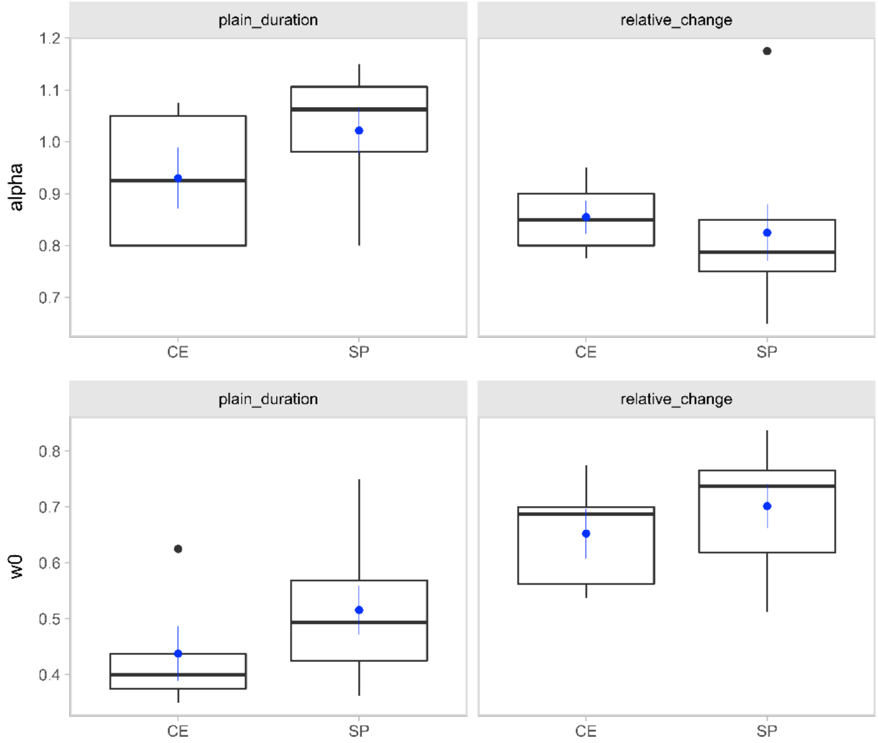

Figure 5. Boxplots of α and w_0 values for the *plain duration* and *relative change* methods, and for CE and SP speakers. Blue dots and blue lines are the means and standard errors, respectively.

In order to assess the tendencies revealed by the descriptive statistics and the boxplots presented as well as to evaluate possible effects of dialect and method on α and w_0, two Bayesian mixed-effects regression models were fitted, one to the α data and another to the w_0 data. Both models include random intercepts for “speakers”, since each contributed with more than one data point. We also fitted a model with random slopes for "dialect", and another one with an interaction between dialect and method (which was not significant), but a model comparison using the Bayesian leave-one-out cross-validation[9] indicated the model with only varying intercepts with no interaction as the the one with more predictive ability, both for α and for w_0 .

[9] Using the LOO() function from the brms package for R (BÜRKNER, 2017; 2018).

Since there is not enough previous literature to justify expected values of α *a priori* for Brazilian Portuguese, flat priors[10] were used for the α model, whose coefficients are presented in Table 2.

	alpha		
Predictors	*Estimates*	*CI (50%)*	*CI (95%)*
Intercept	0.97	0.93 – 1.01	0.85 – 1.09
method: relative_change	-0.15	-0.18 – -0.12	-0.24 – -0.06
dialect: SP	0.03	-0.02 – 0.08	-0.12 – 0.17
Random Effects			
sd(Intercept)	0.08		0.00 – 0.18
Observations	26		
Marginal R^2 / Conditional R^2	0.298 / 0.530		

Table 2. Coefficients of the Bayesian mixed-effects model fitted to the α data. Model: alpha ~ method + dialect + (1 | id)

In a Bayesian model, coefficients are given as probability distributions, and not as point estimates. Therefore, the values under "Estimates" in table 2 are simply one of the most probable values of each parameter (median of the distribution), but any value within the central portion of the distribution is very likely. This is why table 2 also presents the intervals of the 50% and the 95% most credible values, for comparison and assistance in interpreting the model. The estimate for the "Intercept" is that for the first level of each predictor variable in alphabetical order, so the 0.97 for the intercept indicates the most probable value of α for CE speakers under the *plain duration* method (but ranging from 0.85 to 1.09 in its 95% most credible interval). In the second row we can see what happens to this probable α value when changing the method from *plain duration* to *relative change* – it decreases α by -0.15 (it could be from as much as -0.24 to just -0.06, according to its 95% credible interval, but still meaning a decrease). Changing the dialect from CE to SP (third row) might seem to increase α values if one only looks at the most probable value for this estimate, but the negative values within the 95% credible interval (and even within the 50% CI) show that the model cannot really tell if this change in dialect increases or decreases α. The random

[10] The default values given by the *brms* package for R (BÜRKNER, 2017; 2018).

effects showed a probable standard deviation of 0.08 for individuals, meaning we could expect 68% of participants to vary their intercepts in + or - 0.08 (1 SD), and 95% to vary their intercepts in + or - 0.16 (2 SDs)[11].

Figure 5 presents the posterior probability distributions of the coefficients of the model, with the thick blue lines at the medians of the distributions (the "Estimates" from table 2), the shaded blue areas corresponding to the 50% CIs, and the tails of each distribution bounded at its 99% CI.

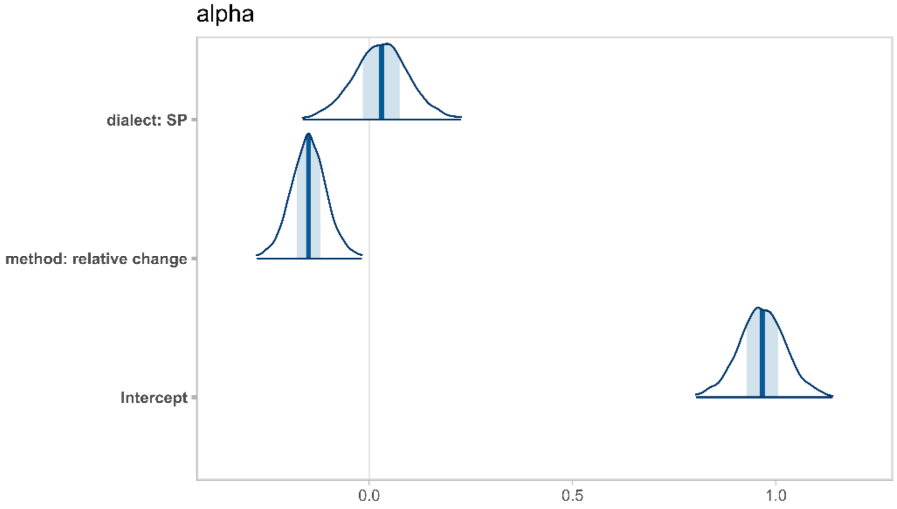

Figure 6. Posterior probability distributions of model coefficients for α.

The third distribution shows the most probable values of α for CE speakers under the *plain duration* method (Intercept). The second distribution, completely on the negative side, makes us confident that using the *relative change* method decreases α values. The first distribution, on the other hand, with 34% of its area on the negative side, indicates that changing the dialect from CE to SP, at least with our data, does not affect α[12].

The graph with fitted values, in figure 7, reinforces the lack of effect of dialect in our data and the decrease of α values in the *relative change* method. Keeping the same system used so far, the dots represent the most probable value (median of the distribution) of α, the thicker portion of each line represents its 50% most credible interval, and the thinner portions extend up to the 95% CI.

[11] Note, though, that random effects are also given as probability distributions and the 95% credible interval of the effect is given in the table.

[12] Having 66% of the area of the posterior distribution on the positive side is not enough to recognize an effect of dialect. If we took the 95% / 5% (e.g., p = 0.05) value commonly used in frequentist statistics for decision making, 66% would be a definite "no effect".

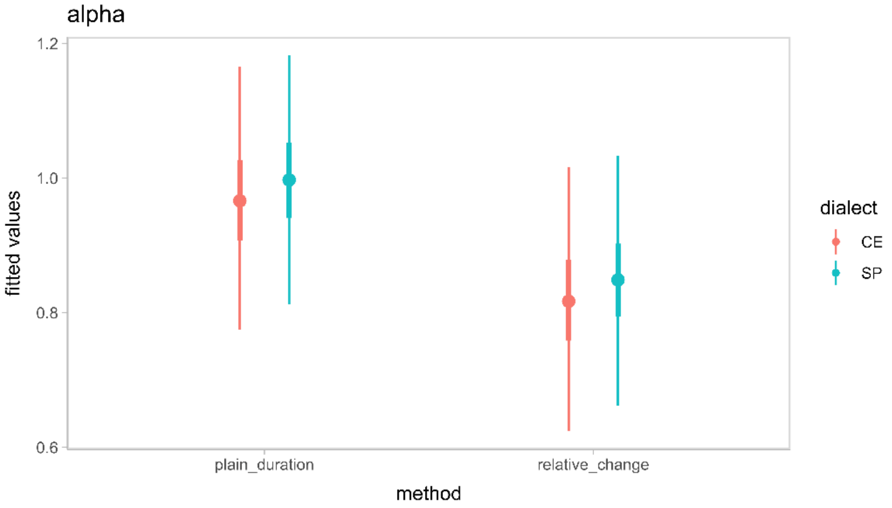

Figure 7. Fitted values of α for CE and SP dialects, and for *plain duration* and *relative change* methods.

Assuming Brazilian Portuguese as having a mixed rhythm, we set a regularizing prior distribution for w_0 centered at 0.5, but allowing for values between 0 and 1. We defined the prior for the Intercept as a normal distribution with mean 0.5 and standard deviation 0.25. The prior for the slopes were also regularizing priors, centered at zero to allow for changes in any direction (increasing or decreasing w_0). It was defined as a normal distribution with mean 0 with a standard deviation of 0.35, to allow for both the increase or decrease of w_0, but within the 0 to 1 range. Finally, the prior for sigma was set as a normal distribution with mean 0 and standard deviation 0.25 (truncated at 0).

The posterior distributions of the coefficients of the model for w_0 are presented in table 3, with the same style previously used to report on α.

	w0		
Predictors	*Estimates*	*CI (50%)*	*CI (95%)*
Intercept	0.45	0.42 – 0.48	0.35 – 0.54
method: relative_change	0.19	0.17 – 0.22	0.11 – 0.28
dialect: SP	0.06	0.03 – 0.10	-0.05 – 0.17
Random Effects			
sd(Intercept)	0.05		0.00 – 0.12
Observations	26		
Marginal R^2 / Conditional R^2	0.487 / 0.555		

Table 3. Coefficients of the Bayesian mixed-effects model fitted to the w_0 data. Model: w0 ~ method + dialect + (1 | id)

The most probable value for w_0 for a CE speaker in the *plain duration* method (Intercept) is 0.45 (ranging from 0.36 to 0.55 in the 95% most credible interval). Changing the method to *relative duration* increases w_0, the most probable value being +0.2, but ranging from 0.11 to 0.28 in the 95% credible interval. Changing the dialect to SP seems to increase w_0 values, but we cannot be certain at this point because part of the 95% CI of this effect crosses zero, showing there is probability that there is no effect of dialect. The 50% most credible interval stays in the positive side, showing an increase of w_0 of around 0.03 to 0.1, which is probably still very low to have any linguistic effect on one's perceived rhythm of speech. At the present time, there is no objective way of establishing the linguistic significance of differences in w_0 for rhythmic classification purposes either on production or perceptual grounds. This should be a topic for future research. The random effects show a probable standard deviation of 0.05, which represents the variation of individual values in the intercept. Individual variation around w_0 was slightly smaller than individual variation in α.

Figure 8 presents the posterior probability distributions of the coefficients of this model. Just as was done with the previous model, the thick lines are the medians of the distributions, the shaded blue areas correspond to the 50% most credible interval of each distribution, and the tails of the distributions are bounded to the 99% CI.

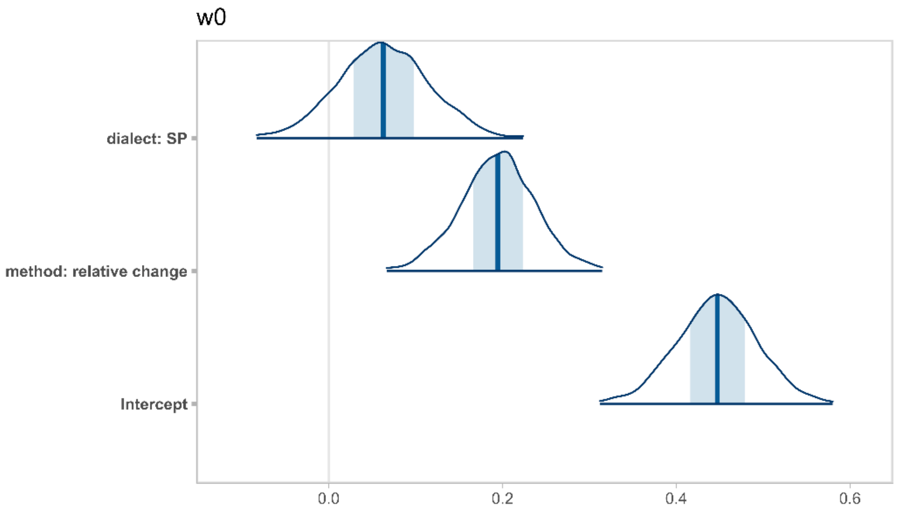

Figure 8. Posterior probability distributions of model coefficients for w_0.

The bottom distribution shows the most probable value of w_0 for a speaker from CE under the *plain duration* method (Intercept from the model). The distribution in the middle, completely on the positive side, indicates that the *relative change* method yielded higher values for w_0. The distribution on the top has 13% of its area on the negative side, making us skeptical about an effect of dialect on w_0 with our data.

The graph with fitted values, in figure 9, reinforces the increase in w_0 values derived by the *relative change* method. It also shows that the difference between CE and SP is higher for w_0 than it was for α (figure 7), but still not high enough to lead to the conclusion of an effect of dialect in our data. The graph in figure 9 uses the same system used so far, where the dots represent the most probable value (medians of the distributions) of w_0, the thicker portion of each line represents its 50% most credible interval, and the thinner portions extend up to the 95% CI.

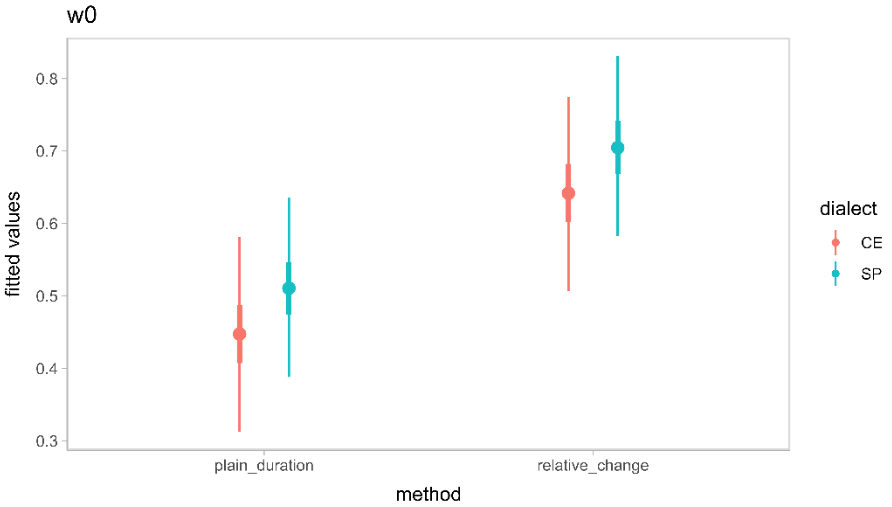

Figure 9. Fitted values of w_0 for CE and SP dialects, and for *plain duration* and *relative change* methods.

3. DISCUSSION

From the point of view of the development of our methodology, the results are relevant because they show that the contour comparison method has an important effect on the estimates of both w_0 and α values. The *relative change* method moves the w_0 estimates up by 0.19 in comparison with *plain duration*, while lowering α by 0.15. In other words, the changes caused by the methods go in opposite directions for w_0 and α, almost as if to cancel each other out. Since both differences are credible and their effect is not negligible, given that w_0 varies between 0 and 1, further analyses of these results are necessary, so a principled decision can be made about which method to choose as the default for future work.

From the linguistic point of view, even if the results are affected by the comparison method, they show relevant trends. Regarding the effect of dialect, the present results do not show compelling evidence of differences either in w_0 or in α when we look separately at contour comparison methods. There is also not strong evidence in favor of a difference between w_0 and α in terms of variability. These results seem to confirm some of the previously held broad assumptions about BP's rhythmic tendencies. Results also agree with Barbosa's assumption that w_0 values should be fairly similar among varieties of the same language.

As for α, the results show that both varieties have similar values as well, and its overall variability is only slightly higher than the one seen for w_0, not enough to allow us to infer that it has greater variability These results do not necessarily contradict Barbosa's assumptions, because the author does not assume that α *should* be more variable than w_0. Our sample is too small in order to compare between-speaker variability with between-dialect variability, so a larger sample would unveil a difference in α if there is one. Given the scarcity of previous studies on rhythmic differences between the various BP varieties, we cannot yet firmly assume that the lack of dialect effect reflects the true nature of the two varieties or is a by-product of our procedure.

4. FINAL REMARKS

Given the complex nature of speech rhythm and the number of degrees of freedom involved in the design of a procedure such as the one we are trying to develop, the results presented here are encouraging. Except for the labeling part, the procedure is driven by scripts, making it easier to scale up studies on rhythmic characterization and variability. In comparison with Barbosa's procedure (BARBOSA, 2004), it has the advantage of also estimating α and β.

Regarding the development of the procedure, the present results indicate that further analysis of the performance of the contour comparison methods presented here are needed to gain a better understanding of the differences and make an informed decision on which of the two should be used going forward.

Notwithstanding the pending issue of the contour comparison methods, the present results suggest that the procedure generates linguistically sensible estimates for w_0 and α, corroborating the basic assumption that BP is a mix-type language in terms of rhythmic organization. The results also lean towards corroborating some assumptions made by Barbosa about the cross-dialectal and between-speaker variability of w_0 and α, although firmer conclusions require a larger sample.

On the linguistic side, once it becomes possible to select one single contour comparison method, we plan to apply the methodology to languages other than BP. Crucially, we plan to look at languages that are widely regarded in the literature as prime examples of stress-timed and syllable-timed rhythm, such as English and Spanish, respectively. This will allow us to test some of the most crucial hypotheses raised by Barbosa regarding the role of the coupling strength parameter (w_0) as an index of rhythm typology. Further in the future, if the methodology proves to be sound, we also plan to apply it to the speech of L2 learners.

5. ACKNOWLEDGEMENTS

The first author would like to thank Plinio A. Barbosa for sharing the sound files for the São Paulo variety analyzed here. The second author would like to thank the Brazilian National Council for Scientific and Technological Development – CNPq (process 438823/2018-4) for partially supporting this project. The authors also thank the reviewers for their careful reading and for their comments, which greatly improved the paper.

REFERENCES

ABERCROMBIE, David. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press, 1967.

BARBOSA, P. A. Elementos para uma tipologia do ritmo (lingüístico) da fala à luz de um modelo de osciladores acoplados. *In Cognito - Cadernos Românicos em Ciências Cognitivas*, v. 2, n. 1, p. 31–58, 2004.

BARBOSA, P. A. Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. 2002, Aix-en-Provence, France. *Anais*... Aix-en-Provence, France: [s.n.], 2002. p. 163–166.

BARBOSA, P. A. From syntax to acoustic duration: a dynamical model of speech rhythm production. *Speech Communication*, v. 49, p. 725–742, 2007.

BARBOSA, P. A. *Incursões em torno do ritmo da fala*. Campinas: Pontes, 2006.

BERTINETTO, Pier Marco. Reflections on the dichotomy "stress" vs. "syllable timing". *Révue de Phonétique Appliquée*, v. 91, p. 99–129, 1989.

BÜRKNER, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1–28, 2017. doi: 10.18637/jss.v080.i01.

BÜRKNER, P. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, 10(1), 395–411, 2018. doi: 10.32614/RJ-2018-017.

CELCE-MURCIA, M. *et al. Teaching Pronunciation: a course book and reference guide*. Cambridge, UK: Cambridge University Press, 2010.

DAUER, R. M. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, v. 11, p. 51–62, 1983.

LADEFOGED, Peter. *A Course in Phonetics*. 4th. ed. Boston, MA: Heinle & Heinle, 2001.

PETTORINO, Massimo *et al.* VtoV: a perceptual cue for rhythm identification. In: PROSODY-DISCOURSE INTERFACE CONFERENCE 2013, 2013, [S.l: s.n.], 2013. p. 101–106.

PIKE, Kenneth Lee. *The Intonation of American English*. Ann Arbor: University of Michigan Press, 1945.

APPENDIX

The Celce-Murcia (2010) Passage

"O inglês é a sua língua nativa? Caso não seja, o seu sotaque estrangeiro pode mostrar para as pessoas que você vem de outro país. Por que é difícil falar uma língua estrangeira sem sotaque? Existem algumas respostas para essa pergunta. Primeiro, a idade é um fator importante na aprendizagem da pronúncia. Nós sabemos que crianças pequenas conseguem aprender uma segunda língua com pronúncia perfeita. Também sabemos que aprendizes mais velhos normalmente têm sotaque, apesar de alguns aprendizes mais velhos também conseguirem aprender a falar sem sotaque algum.

Outro fator que influencia a pronúncia é a sua língua materna. Falantes de inglês conseguem, por exemplo, reconhecer franceses por seus sotaques franceses. Eles também conseguem identificar falantes de espanhol ou de árabe ao telefone, apenas por escutá-los com cuidado. Isso significa que sotaques não podem ser mudados? De maneira nenhuma! Mas você não consegue mudar a sua pronúncia sem trabalhar bastante nisso. Afinal, melhorar a pronúncia é uma combinação de três coisas: muito trabalho concentrado, um bom ouvido, e uma forte ambição de soar como um falante nativo.

Você também precisa de informações precisas sobre os sons do inglês, estratégias eficientes para praticar, muita exposição ao inglês falado, e paciência. Você fará progresso ou desistirá? Apenas o tempo dirá. Mas é sua decisão. Você pode melhorar! Boa sorte, e não se esqueça de estudar e praticar bastante!"

The Lobato Passage

"Em seguida apareceu um papagaio real que tinha fama de orador. Subiu a tribuna de um poleiro de ouro e fez um belo discurso a respeito da arte de falar. Nesse discurso provou que os homens tinham aprendido a falar com os papagaios, e não os papagaios com os homens, como diz a ciência destes. Uma chuva de palmas acolheu suas palavras.

O mesmo não aconteceu, porém, com a poetisa Lagartixa, que principiou a recitar uma longa poesia e engasgou no meio, acabando o recitativo em choro e faniquito. Para destruir essa má impressão vieram três vagalumes mágicos que fizeram várias sortes, sendo muito apreciada a sorte de comer fogo."

Lobato, M. (1920) A Menina do Narizinho Arrebitado. São Paulo: Revista do Brasil. Monteiro Lobato & Cia. p. 18 (ortografia modernizada).