Annotation–mining for rhythm model comparison in Brazilian Portuguese

Dafydd Gibbon[1], *Flaviane Romani Fernandes*[2]

[1]Universität Bielefeld, Germany, [2]Universidade Estadual de Campinas, Brazil
[1]gibbon@uni-bielefeld.de, [2]flaviane@gmail.com

Abstract

Speech rhythm has been investigated from phonetic, phonological and signal processing perspectives, leading to a wide range of non–comparable methodologies. We assume that this is because speech rhythm is an emergent phenomenon (Emergent Rhythm Theory), due to many 'hidden' physiological and other factors, and that specialists make different selections from these factors. We also claim that models proposed so far are relatively arbitrary, and that their formal and empirical similarities and differences are not clarified. Although we accept Emergent Rhythm Theory, we acknowledge that it is currently too complex and inexplicit to be falsifiable, and concentrate on more highly constrained Physical Rhythm Theory approaches. We propose a set of explicanda for Physical Rhythm Theories, and formally and empirically compare selected single–parameter physical rhythm measures. We find that the selected measures show a very low degree of similarity and outline the steps necessary to improve the situation.

1. Introduction

In the present study we note that speech rhythm has been investigated from many perspectives, including those of phonetics, phonology, signal processing and discourse analysis, with a wide range of non–comparable methodologies and divergent results. We claim that this is partly due to to the character of speech rhythm as an emergent phenomenon due to many 'hidden' factors, not just physical (Emergent Rhythm Theory, ERT), that specialists make different selections from these factors, that the modelling conventions used in the literature are not sufficiently explicit, that formal and empirical similarities and differences between rhythm models remain unclear, and that measures proposed so far are relatively arbitrary and unrelated. We therefore compare selected rhythm models empirically, starting with basic single–parameter Physical Rhythm Theories (PRT) [1], [2], [3], and outline a research programme for comparison of stepwise more complex theories, using a corpus–linguistic technique which we term 'annotation–mining'. Section 2 briefly discusses the background of ERT and PRT, and introduces explicit modelling conventions for rhythm. Section 3 outlines the basic formal properties of a selection of PRT models and reports on an empirical comparison of the selected models. Section 4 discusses the results, followed by the conclusion and outlook in Section 5.

2. Physical cues and emergent rhythm

Our goal is to compare models of rhythm with the aim of finding formal and empirical evaluation criteria for these models. In the first instance we are do not ask whether a model is 'correct' or 'better' than other models; we just aim to examine the formal and empirical similarity of models.

2.1. Constraining rhythm models

For this purpose it is necessary to constrain views on rhythm. Fundamentally, we hold the ERT view that rhythm is an emergent perceptual construct based on the coordination of many different temporal activities due to the interaction of a variety of different physiological and cognitive systems. A simple analogy is found in 'beat' or 'heterodyne' frequencies, ('Tartini overtones' in violin technique): a third output frequency is a function of the non–linear mixing of two input frequencies. Similarly, linguistic approaches to rhythmic organisation, particularly Metrical Phonology, combine syntactic structure, readjustment rules and grid constraints, potentially also semantic and pragmatic focus and discourse constraints: $rhythm = f(discourse, focus, syn, readj, grid, phon)$.

But because of its complexity and, overall, its inexplicitness, we also assume that ERT is unlikely to become a falsifiable theory in the near future, though entrainment and oscillation theories [4], [5] show promise in this area. We therefore constrain this approach by heuristically adopting the PRT standpoint that there are indeed physical cues to rhythm (by no means a necessary assumption):

1. The signal provides cues for synchronising with the constrained activities which produced it.
2. Cues to rhythmical organisation can be detected by distributional analysis of physical measurements.
3. But: careful subjective annotation approximates to a criterion for emergent phenomena.

Even so, rhythm may also be an emergent function of physical cues for prominence/nonprominence alternation, such as syllable or foot structure, or a sonority criterion [6]. Consequently we need still further constraints. We therefore lower our sights again and examine PRT approaches in respect of formal and empirical similarity, on the hypothesis that at least the PRT models of rhythm will turn out to be formally and empirically quite similar. Rather than taking the line of [7], examining subsyllabic units with a 2–parameter model, we concentrate on 1–parameter models applied to syllable structure alone.

The longer–term strategy is gradually to relax the constraints in a well–defined manner and define a model space in which approaches to rhythm modelling are located. The results are intended to be of direct use in the formal and empirical modelling of timing, for theoretical purposes in linguistics and phonetics, but also for applications such as speech synthesis.

2.2. Modelling conventions for rhythm

Several recent PRT oriented studies have produced interesting results for prosodic typology by applying different physical measures to temporally annotated signal data, in particular a 2–parameter [8], [7] and a 1–parameter [3] approach. Our interest is also in comparing prosodic variation, both inter–language

Table 1: Formal comparision of Linear Rhythm Models with reference to rhythm modelling conventions.

Name	Model	*Base unit*	*Alternation*	*Iteration*	*Isochrony*
PIM	$\Sigma_{i \neq j} \left\lvert \log \frac{I_i}{I_j} \right\rvert$	foot	no	no	yes
PFD	$100 \times \frac{\Sigma \lvert MFL - len(foot_i) \rvert}{n \times MFL}$, MFL = $\frac{\Sigma_{i=1}^{n} \lvert foot_i \rvert}{n}$	foot	no	no	yes
PVI	$100 \times \Sigma_{k=1}^{m-1} \left\lvert \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right\rvert / (m-1)$	V stretches	no	yes	yes

and intra–language, but in view of the innumerable degrees of freedom in this task we start our model comparisons with a corpus of highly constrained data [9] (though not as highly constrained as that of [5]). In addition we introduce an explicit set of modelling conventions for developing and evaluating PRT approaches:

Base unit: a finite trajectory through an n-dimensional parameter space (pitch, segment, syllable, foot sequence...).

Alternation: a dynamic, not flat base pattern trajectory, i.e. traversal through at least two positions in the parameter space (varying pitch pattern, CV syllable pattern, long–short or strong–weak syllable foot pattern,...).

Iteration the base pattern P must repeat with at least two occurrences: $P\ P^{+}$, i.e. any of $\{< P_1, P_2 >, < P_1, P_2, P_3 >, ...\}$.

Isochrony : equal length of the base pattern, i.e. $|P_i| = |P_{i+1}|$.

On this basis we examine a set of 1–parameter linear (non–hierarchical) PRT measures of rhythm based on single measures such as syllable or foot duration: two Global Linear Rhythm Models (with no concept of local alternation and iteration), the Pairwise Irregularity Measure (PIM) of [2], and the Percentage Foot Deviation (PFD) model of [1] based on Mean Foot Length (MFL), and one Local Linear Rhythm Model (with a concept of local alternation and iteration), the Pairwise Variability Index (PVI) of [3]. In [10], [11] it was found that the models are all incomplete as PRT models; we summarise this in Figure 1 in terms of the modelling conventions introduced here. Effectively, the models all turn out to be simply isochrony models and the alternation and iteration criteria are not met.

The Rhythm Ratio (RR) of [12] is omitted because it is known to be very similar to the PVI; the $\Delta C \times \%V$ model of [8] and the $\Delta C \times \Delta V$ model of [7] are not investigated at this stage because they operate in a 2–parameter space (C vs. V) and are therefore not directly comparable. Hierarchical models are also not included since suitable quantitative measures which would provide a basis for the comparison are not defined.

3. Comparison of Linear Rhythm Models

In this section we compare Linear Rhythm models using corpus–based empirical criteria [2], [1], [3]. To our knowledge, no direct empirical comparison of these Linear Rhythm Models is available.

3.1. The Brazilian Portuguese corpus

Rhythm and related issues in Brazilian Portuguese have been studied relatively extensively (cf. [4], [13], [14] [15] [16]). These studies consider the iterative and isochrony modelling conventions of rhythm, without reference to other physical, structural, semantic or pragmatic factors. The study by [9], is a phonological study, and considers relevant and functional but not not phonetic factors.

Part of the corpus due to [9] is used, consisting of sentences read aloud. We deliberately chose this data type in order to provide a 'best case' with a highly constrained type of data, a similar strategy, in general terms, to that of [5], but with a less artificial variety of speech.

The corpus is based on readings of 49 sentences that present different syntactic structures: SV, SSV, SVO, SSVO, SSVOO, SVOO, SSVAO, SS and SS (Scomp) V, where S = Subject, V = Verb, O = Object, A = Adverb, SS = Simple Subject constituted by two elements (a substantive and an adjective, for instance 'o sofá preto'—'the black sofa'), OO = an Only Object constituted by two elements, SS and SS (Scomp) = complex subject constituted by four elements (two substantives and two adjectives, for example 'o tatu russo e a abelha rainha'—'the Russian armadillo and the queen bee'). The sentences contain transitive verbs, unaccusative verbs and unergative verbs, under non–focus (neutral) conditions.

The sentences were produced by five educated (graduate) female Brazilian speakers (age range 26 to 51 years) from different regions of Brazil: Rio de Janeiro city (Rio de Janeiro State), São José do Rio Preto city (São Paulo State), Garça city (São Paulo State) and Itumbira city (Goiás State).

3.2. Method

In order to perform these comparisons on the Brazilian Portuguese data as efficiently as possible (242 syllable–annotated files in Praat notation), a suite of Unix scripts was implemented with phonetic components (PIM, PFD, PVI) and statistical components (mean, standard deviation, average deviation, correlation).

In addition, we investigated the issue of variation in temporal rate in the corpus, which was the original motivation for the pairwise comparison criterion of the PVI [3] and the RR [12]. For this purpose we applied least squares regression analysis individually to duration sequences in the sentence annotations, starting with linear regression and then progressing if necessary to more complex functions (in view of the fact that some of the rhythm models are non–linear).

The following procedure was used:

Preprocessing: Extraction of syllable tier durations from Praat annotation file.

Phonetic processing: Application of PIM, PFD and PVI algorithms to duration sequences, separately for all five speakers.

Single speaker analysis: Calculation of mean, standard deviation, average deviation, for phonetic algorithm output.

Regression analysis: Calculation of slope of duration sequences for each speaker.

All speakers analysis: Calculation of overall values for all speakers.

Table 2: Empirical comparison of Linear Rhythm Models.

Speaker	mean	sd	av-dev	PIM	PFD	PVI	intercept	slope
Bre	0.216512	0.0972989	0.0721281	4.52341	33.6465	50.0578	0.394149	0.00293441
MC	0.154891	0.0775258	0.0555343	4.52436	35.6848	46.4656	0.0686522	0.0203175
MR	0.211479	0.097132	0.0720969	4.09595	34.1894	45.3266	0.102436	0.02684
Sil	0.220721	0.106279	0.0800285	4.50612	36.6584	46.9615	0.0977527	0.0302151
Sim	0.201764	0.109419	0.0796148	5.39055	39.5387	51.8874	0.0663943	0.0333393

4. Results and discussion

Table 2 shows the mean duration values for speaker and each measure. The table shows that all the measures are in general comparable across speakers. Exceptions are

- the mean duration for speaker MC, whose speech rate appears to be rather fast, if our data are accurate;
- the linear regression results for speaker Bre: intercept (offset constant) and slope differ from the other values by orders of magnitude.

In order to explain these differences, the signal and annotation data need closer inspection than was possible for this study.

We interpret the extremely low slope values as indicating that the speech tempo does not vary noticeably over the utterances, which is not particularly surprising, since the utterances are rather short single sentences. We conclude that the PVI (and RR) local normalisation for speech tempo rate is unnecessary for isochrony studies of highly constrained data of the reading–aloud type.

Table 3 shows the correlations between each of the measures. It is striking that in most cases the correlation, either positive or negative, is quite low. The best correlations are, unsurprisingly, between the standard deviation and the average deviation, which are formally very similar measures; both (like variance) have been used in other studies, but only one is needed. Otherwise the 'best' correlations are found between the average deviation and the PFD, then between mean and the PIM, and then between the PFD and the PVI. For average deviation, PFD and PVI, the similarities are presumably due to formal similarities between the measures; how far this is true for the other measures requires further analysis.

5. Conclusion and outlook

We defined a set of explicanda for rhythm models and evaluated a set of PRT measures of rhythm (more accurately: of isochrony) with respect to these explicanda. We operationalised these models, applied them to Portuguese Brazilian data, and obtained quantitative rhythm measures for each speaker and model. In order to obtain a 'best case' result, we used a fairly highly constrained corpus of read–aloud, systematically designed sentences. The measures and their correlations for each speaker and method were correlated. The procedure was performed automatically, enabling a reasonably large amount of processing to be performed efficiently. As already noted, the methodology of this annotation mining study is not designed to demonstrate whether one of the studies is 'true', or 'better' than one of the others, but to provide similarity criteria for a comparison of the models relative to each other.

Our expectation was that since all studies claim to be investigating the isochrony dimension of rhythm, the results would at least be in some way rather comparable. The correlation results show, however, that the measures are only weakly similar. There are large margins of unexplained variation. Since the empirical variable, syllable duration, was held constant, and behaviour was rather similar across speakers, it could be expected that the differences may be explained, basically, by the formal properties of the models, some of which are non–linear in different ways.

Other studies examine duration properties of other units such as the foot, or consonantal and vocalic syllable constituents. The choice of alternation unit is not particularly important for this evaluation study, as our intention was to reduce the number of empirical variables involved at this stage, and we picked the syllable alone as a strategy for minimizing the number of variables as a first step, increasing complexity later. As a basic alternation unit for the investigation of rhythm, it is of course likely that the bare syllable is not always the best candidate; consonantal and vocalic sub–syllabic patterns [8], [7], and the foot [1] also need to be examined.

We have established a basic schedule for comparing rhythm models on the basis of their explicanda, their formal properties, and their empirical results. In order to explain the unexpected result that existing isochrony based rhythm models are only weakly similar, further work is needed. We characterise the required extended schedule as follows:

1. extension of the corpus;
2. application of the technique to other corpora, in particular to other languages;
3. investigation of 2–parameter (and n–parameter) Global Linear Rhythm models such as [8], [7];
4. development of quantitative criteria for formally and empirically comparing hierarchical models using the annotation–mining method [10], [11];
5. examination of other alternation units such as consonantal and vocalic sequences below the syllable level, foot sequences above the syllable level, and the interactions between these levels.

Currently we are prioritising items 1, 3 and 5 and concentrating specifically on the definition of a valid model of alternation, taking the oscillation and entrainment models of [4], [13], [14], [5] and [17] into account.

We believe strongly that progress will be furthered best by understanding the relations between rhythm models of various kinds within the formal and empirical model space within which they are located, and that the dimensionality of comparisons within this model space needs to increase step by step in the manner discussed in this contribution.

6. Acknowledgements

This work was partly financed by the projects "Data-mining in large speech corpora" (DAAD PROBRAL programme), and "Ordem e preencimento em português: sintaxe entoação e ritmo" (FAPESP 03/13938-5), and by a Ph.D. sandwich scholarship (CAPES BEX 0183/05-9).

Table 3: Correlation of measures for each speaker.

Speaker	function	sd	av-dev	PIM	PFD	PVI
Bre	mean	0.0748831	0.118778	-0.846811	-0.42066	-0.267731
	sd		0.907009	0.205994	0.77598	0.62168
	avdev			0.143908	0.840816	0.676107
	PIM				0.600022	0.42216
	PFD					0.752826
MC	mean	0.450368	.508914	-0.188575	0.235218	0.0711187
	sd		0.917107	0.546764	0.890715	0.48838
	avdev			0.551634	0.950961	0.535372
	PIM				0.709551	0.542667
	PFD					0.604984
MR	mean	0.499709	0.514251	-0.776711	-0.153547	-0.0374781
	sd		0.90425	-0.162971	0.6622	0.411336
	avdev			-0.223698	0.76285	0.400339
	PIM				0.32755	0.296824
	PFD					0.5007
Sil	mean	-0.0146212	0.0264243	-0.683822	-0.47013	-0.300565
	sd		0.871596	0.134199	0.775744	0.443922
	avdev			0.133069	0.861513	0.454513
	PIM				0.486842	0.473204
	PFD					0.572758
Sim	mean	0.469375	.570693	-0.727528	-0.127606	-0.123911
	sd		0.827847	-0.224215	0.61132	0.239266
	avdev			-0.256116	0.737965	0.285564
	PIM				0.291293	0.287871
	PFD					0.438312

We are especially indebted to Charlotte Galves, Antonio Galves, Filomena Sândalo and Maria–Clara Paixão de Sousa for many discussions on the topics dealt with in this study.

The Unix scripts designed and implemented for this study are available from the authors.

7. References

[1] P. Roach, "On the distinction between 'stress-timed' and 'syllable-timed' languages," in *Linguistic Controversies: Essays in Linguistic Theory and Practice*, D. Crystal, Ed. London: Edward Arnold, 1982, pp. 73–79.

[2] D. R. Scott, S. D. Isard, and B. de Boysson-Bardies, "On the measurement of rhythmic irregularity: a reply to Benguerel," *Journal of Phonetics*, vol. 14, pp. 327–330, 1986.

[3] E. L. Low, E. Grabe, and F. Nolan, "Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English," *Language and Speech*, vol. 43, no. 4, pp. 377–401, 2000.

[4] P. A. Barbosa, "Explaining brazilian portuguese resistance to stress shift with a coupled–oscillator model of speech rhythm production," *Cadernos de Estudos Lingüisticos*, vol. 43, pp. 71–92, 2002.

[5] F. Cummins, "Speech rhythm and rhythmic taxonomy," in *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 2002, pp. 121–126.

[6] A. Galves, J. García, D. Duarte, and C. Galves, "Sonority as a basis for rhythmic class discrimination," in *Speech Prosody 2002*, Aix–en–Provence, 2002.

[7] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," in *Proceedings of Speech Prosody 2002*, Aix–en–Provence, 2002, pp. 115–120.

[8] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[9] M. F. Sândalo and H. Truckenbrodt, "Some notes on phonological phrasing in Brazilian Portuguese," *MIT Working Papers in Linguistics*, vol. 42, 2002.

[10] D. Gibbon, "Computational modelling of rhythm as alternation, iteration and hierarchy," in *Proceedings of ICPhS 2003*, Barcelona, 2003.

[11] ——, "Corpus–based syntax-prosody tree matching," in *Proceedings of EUROSPEECH 2003*, Geneva, 2003.

[12] D. Gibbon and U. Gut, "Measuring speech rhythm in varieties of English," in *Proceedings of EUROSPEECH 2001*, Aalborg, 2001, pp. 91–94.

[13] P. A. Barbosa, "Integrating gestural temporal constraints in a model of speech rhythm production," *Journal of Phonetics, Special issue: Temporal Integration in the Perception of Speech, ed. Sarah Hawkins & Noël Nguyen*, vol. 31 3/4, p. 54.

[14] ——, "Explaining cross-linguistic rhythmic variability via a coupledoscillator model of rhythm production," in *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 2002, pp. 163–166.

[15] D. Duarte, A. Galves, N. Lopes, and R. Maronna, "The statistical analysis of acoustic correlates of speech rhythm. The Sciences of Complexity Conference, ZiF, Bielefeld," 2001.

[16] S. Frota and M. Vigário, "Aspectos de prosódia comparada: ritmo e entoação no PE e no PB," *Actas do XV Encontro Nacional da Associação Portuguesa de Lingüistica*, vol. 1, pp. 533–555.

[17] I. Wachsmuth, "Communicative rhythm in gesture and speech," in *Language, Vision and Music*, P. McKevitt, C. Mulvihill, and S. Ó'Nualláin, Eds. Amsterdam: John Benjamin, 2002, pp. 117–132.